Expectation-Maximization enables Phylogenetic Dating under a Categorical Rate Model

Uyen Mai[1], Eduardo Charvel[2], and Siavash Mirarab[3*]

[1] *Department of Computer Science and Engineering, UC San Diego, CA 92093, USA*
[2] *Bioinformatics and Systems Biology Graduate Program, UC San Diego, CA 92093, USA*
[3] *Department of Electrical and Computer Engineering, UC San Diego, CA 92093, USA*
[*] Corresponding author: smirarab@ucsd.edu

Abstract

Dating phylogenetic trees to obtain branch lengths in the unit of time is essential for many downstream applications but has remained challenging. Dating requires inferring mutation rates that can change across the tree. While we can assume to have information about a small subset of nodes from the fossil record or sampling times (for fast-evolving organisms), inferring the ages of the other nodes essentially requires extrapolation and interpolation. Assuming a clock model that defines a distribution over rates, we can formulate dating as a constrained maximum likelihood (ML) estimation problem. While ML dating methods exist, their accuracy degrades in the face of model misspecification where the assumed parametric statistical clock model vastly differs from the true distribution. Notably, existing methods tend to assume rigid, often unimodal rate distributions. A second challenge is that the likelihood function involves an integral over the continuous domain of the rates and often leads to difficult non-convex optimization problems. To tackle these two challenges, we propose a new method called Molecular Dating using Categorical-models (MD-Cat). MD-Cat uses a categorical model of rates inspired by non-parametric statistics and can approximate a large family of models by discretizing the rate distribution into k categories. Under this model, we can use the Expectation-Maximization (EM) algorithm to co-estimate rate categories and branch lengths in the time unit. Our model has fewer assumptions about the true clock model than parametric models such as Gamma or LogNormal distribution. Our results on two simulated and real datasets of Angiosperms and HIV and a wide selection of rate distributions show that MD-Cat is often more accurate than the alternatives, especially on datasets with nonmodal or multimodal clock models.

Code availability. The LogDate code is available on https://github.com/uym2/MD-Cat.

Data availability. All the data are available on https://github.com/uym2/MD-Cat-paper.

1 Introduction

The rates of mutation can and do change through time, creating a major challenge for dating phylogenetic trees. Since time and mutation rates are inseparable from sequence data, dating relies on pre-specified ancestral divergence times (e.g., calibration points obtained from carbon dating fossils) or sampling times of the tips (available for phylodynamic data) to estimates the rates of evolution and date the tree. However, even with such external information, extrapolating from some known times to all tree branches essentially requires assumptions about how the rates may or may not have changed. Ideally, we need to know a *molecular clock* model, giving us a way to assign probabilities to mutation rates throughout the tree.

The simple strict clock model Zuckerkandl (1962) is insufficient, especially at long evolutionary time horizons where mutation rates have had enough time to change (Bromham and Penny, 2003; Kumar, 2005). To relax the strict clock, uncorrelated and autocorrelated models have been developed. Uncorrelated models assume that rates are drawn i.i.d. for each branch from a parametric distribution, such as exponential, gamma, or lognormal (Aris-Brosou and Yang, 2002), while autocorrelated models allow the rate to evolve on the tree and capture the correlation between parent and child branches. Computational methods for molecular dating have been developed for both uncorrelated and autocorrelated clock models (see Ho and Duchêne, 2014; Kumar and Hedges, 2016; Rutschmann, 2006; Sanderson, 1998). Bayesian methods (Drummond and Rambaut, 2007; Guindon, 2010; Heath et al., 2014; Rannala and Yang, 2007; Thorne and Kishino, 2002; Thorne et al., 1998) can incorporate complex prior models thanks to the use of MCMC sampling methods. These methods allow many types of priors for both molecular clock and calibration points and can infer a dated tree directly from sequence data. However, choosing the model and the prior is challenging, and there is debate around the merit of these Bayesian methods (Beaulieu et al., 2015; Wertheim et al., 2010). Moreover, the computational burden of the MCMC process limits its scalability.

Maximum likelihood (ML) methods also exist (e.g., Langley and Fitch, 1974; To et al., 2016; Volz and Frost, 2017). Assuming a statistical model for the rates, we can seek parameters that maximize the likelihood. ML-based methods usually do not work directly on sequence data but take as input a tree in substitution units (such a tree can be estimated from sequence data using either distance-based (Lefort et al., 2015; Saitou and Nei, 1987) or ML-based methods (Guindon et al., 2010; Nguyen et al., 2015; Stamatakis, 2006)) and convert its branch lengths to time units using calibration points. Although they are more efficient than Bayesian methods and do not need a prior, ML-based methods have their own limitations. The simplest ML methods assume a strict clock model and use either a Poisson (Langley and Fitch, 1974) or Gaussian (To et al., 2016) distribution to model the uncertainties of the estimated branch lengths in substitution units. Such ML methods are fast but restrictive to the strict clock model. More sophisticated ML-based methods, such as TreeDater (Volz and Frost, 2017) and TreeTime (Sagulenko et al., 2018), use a Gamma or Lognormal distribution to model the molecular clock. Because they assume a parametric – and often unimodal – clock model, these ML-based methods may not be robust under complex models with local clocks or sudden rate changes. In addition, the mutation rates of these models are continuous latent variables, so the likelihood function contains an intractable integral, making the likelihood function difficult to optimize. Therefore, some ML-based methods (e.g., Volz and Frost, 2017) depend on heuristic algorithms to iteratively optimize their likelihood functions. These heuristic algorithms lack theoretical supports, such as a guarantee to reach a local (or global) optimum, a monotonic improvement of the likelihood function, or a convergence guarantee.

Recognizing difficulties of relying on a specific clock model, many authors, including us (Mai and Mirarab, 2020), have proposed non-parametric methods for dating (e.g., Sanderson, 1997, 2002; Tamura et al., 2012, 2018; Xia and Yang, 2011). These methods formulate dating as an optimization problem — often in a least-square form — to optimize a predefined objective function without an explicit parametric model of the clock. However, the objective functions used in these methods come from implicit assumptions about the rate distribution. For example, most methods implicitly assume that rates are distributed following a unimodal distribution; that is, the rate of each branch is centering either around a global rate (Mai and Mirarab, 2020; Xia and Yang, 2011) or the rate of its parent or sibling (Britton et al., 2007; Sanderson, 1997; Tamura et al., 2012). Incidentally, minimizing the residual sum-of-square often winds up being the ML solution under specific (often unimodal) models; e.g., our attempt at developing the non-parametric method wLogDate (Mai and Mirarab, 2020) produced a method that is the ML estimate under a set of (unimodal) LogNormal distributions. Thus, although non-parametric methods are fast and show signs of robustness under multiple clock models, we postulate (and test) that these non-parametric methods are inaccurate

when the true clock model is multimodal or is a combination of multiple local clocks.

We make an attempt to combine the best properties of ML and non-parametric methods by introducing a new method, namely Molecular Dating with Categorical rates (MD-Cat). Unlike other ML-based methods, we use a categorical distribution (CAT model) to approximate the unknown continuous clock model (Fig. 1). While categorical models have been adopted to approximate the Gamma model for rate heterogeneity across sequence sites (Felsenstein, 2004; Nguyen et al., 2015; Stamatakis, 2006), the power of the CAT model to approximate an unknown continuous model across branches has not been studied. The CAT model is non-parametric in the sense that it does not assume the rates are drawn from any predefined parametric model. The model is defined as a set of k rate categories at unknown positions from which each rate is drawn. Although the rate categories are discrete, the model has the power to approximate a continuous clock model if k is large and there are enough data to fit the model. We use Expectation-Maximization (EM) algorithm to maximize the likelihood function associated with this model, where the k rate categories and branch lengths in time units are modeled as unknown parameters and co-estimated. We show that both the E-step and M-step of our EM algorithm can be computed efficiently, and the algorithm is guaranteed to converge. We compare MD-Cat to state-of-the-art methods on a diverse set of simulations, including multimodal models and local clocks.

2 Methods

2.1 Notations and Model

For a given binary tree T with n leaves and $N = 2n - 2$ branches, we give each of the $n - 1$ internal nodes of T a unique index in $\{0, \ldots, n-2\}$ (reserving 0 for the root) and give each of the n leaf nodes a unique index in $\{n-1, \ldots, N\}$. We denote the parent of node i as $par(i)$, the left and right children of node i as $c_l(i)$ and $c_r(i)$, and the edge connecting i and $par(i)$ as e_i for $i \in \{1, \ldots, N\}$. The length of e_i is specified in either substitution unit or time unit. Let t_i denote the divergence time of node i (i.e. the time when species i diverges into $c_l(i)$ and $c_r(i)$). Then $\tau_i = t_i - t_{par(i)}$ is the length of e_i in time unit (Fig. 1a). As a shorthand, we combine all the τ_i values into a vector $\tau = [\tau_1 \tau_2 \ldots . \tau_N]$. Similarly, let b_i be the length of e_i in substitution unit, which is the *expected number of substitutions per sequence site* occurred on edge e_i and let $\mathbf{b} = [b_1 b_2 ... b_N]$. We assume that the mutation rate can change only at species divergence times and let $\mu_i = b_i/\tau_i$ denote the rate along branch e_i. Finally, we let s denote the alignment length.

We assume sequences of length s evolve down a binary tree T with mutation rate μ_i along each edge e_i. We assume μ_i are i.i.d following an unknown distribution (e.g., Fig 1b) and sites of the molecular sequences evolve independently following a homogeneous process (such as the GTR model (Tavaré, 1986)). Using Maximum Likelihood (ML) inference, we can obtain an *estimate*, with some uncertainties, of the topology of T and all branch lengths in substitution unit. We assume the tree topology is estimated correctly and model the branch length estimation error using a Gaussian model. Let B_i be a random variable denoting the estimated length of branch e_i in substitution unit and $\hat{b_i}$ be its estimate (Fig. 1a). The distribution of B_i depends on several factors such as the sequence evolution model, sequence length, and the inference technique. For simplicity, we model B_i similarly to To et al. (2016): let $\epsilon_i = B_i - b_i$ be the estimation noise, we assume the noise is Gaussian; that is, $\epsilon_i \sim_{i.i.d.} N(0, \frac{b_i}{s})$, where the variance b_i/s comes from an approximation of the Poisson model of the number of substitutions per sequence site. In addition, for computational and algorithmic convenience, we approximate the variance $\frac{b_i}{s}$ by $\frac{\hat{b_i}}{s}$. Recall that $b_i = \mu_i \tau_i$ and $B_i = \epsilon_i + b_i$, so we have $B_i \sim_{i.i.d.} N(\mu_i \tau_i, \frac{\hat{b_i}}{s})$. Thus, $f(\hat{b_i}|\mu_i, \tau_i)$ is the Gaussian pdf.

As mentioned previously, the clock model that describes the distribution of μ is uncertain. Common choices include Gaussian (Sagulenko et al., 2018; To et al., 2016), Gamma (Volz and Frost, 2017), LogNormal, and Exponential distributions (Drummond et al., 2006). However, there is no guarantee that these distributions, which happen to be unimodal or mode-less, would be good models of how μ changes across a tree. For example, a bimodal or tri-modal set of rates would not be adequately modeled with any of these models. Instead of using a parametric continuous distribution, we use a non-parametric approach using a discrete distribution. We discretize μ into k categories $\omega = [\omega_1, \omega_2, \ldots \omega_k]$ each with the same probability mass $\frac{1}{k}$ and use this categorical model to approximate the unknown distribution that μ was drawn from. Putting together all these elements, we obtain a hierarchical model (Fig. 1c) where τ and ω are parameters,

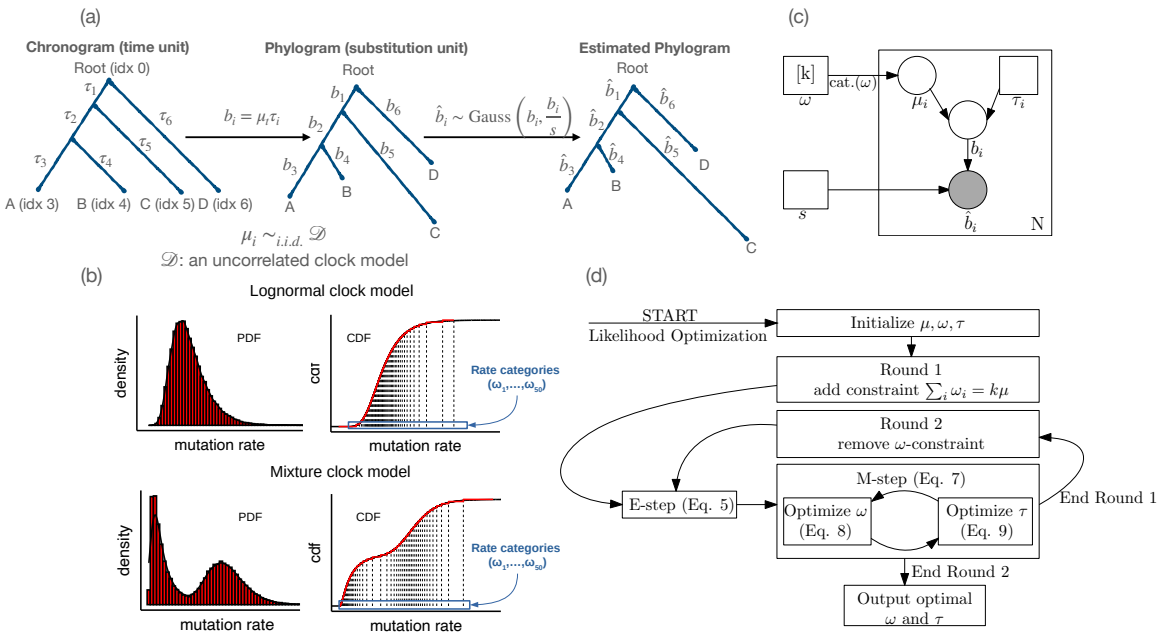

Figure 1: (a) The generative model. (b) Examples of clock models and their categorical approximation. The approximation can be visualized on the PDF as a histogram and on CDF as a step function (shown with $k = 50$). Because we give the same weight to each rate category, the step function has of equal vertical steps. Rate categories are identified by the projection of the steps onto the x-axis of the CDF plot. (c) The hierarchical categorical clock model. We use the plate notation where squares denote parameters, circles denote random variables, and solid circles denote observed data. The number inside a plate (e.g. N) denotes the number of repeats of each random variable inside that plate. (d) Steps of the optimization algorithm.

$\hat{b}_i$'s are observations (i.e. data), and both b_i's and μ_i's are latent variables.

The linear constraints defined by calibration points

Let t_0 be the *unknown* divergence time at the root of the tree. We incorporate the p calibration points $t_1, \ldots, t_p$ as a set of p constraints $C_1, ..., C_p$:

$$C_i : \sum_{j \in P(0,i)} \tau_j = t_i - t_0 \tag{1}$$

where $P(x, y)$ denotes the path between two nodes x and y (referring to by their indices). Recall that the root has index 0, so $P(0, i)$ is the path from the root to node i. To remove t_0 from this set of constraints, we arbitrarily select a constraint C_k and subtract it from the other constraints C_i $(i \neq k)$ side by side to obtain the set $\Psi^{(k)}$ of $p - 1$ linear constraints on τ:

$$\Psi_i^{(k)} : \sum_{j \in P(\mathrm{LCA}(i,k),i)} \tau_j - \sum_{j \in P(\mathrm{LCA}(i,k),k)} \tau_j = t_i - t_k, \tag{2}$$

It is easy to see that all the linear constraint sets $\Psi^{(k)}$ $(k \in [p])$ are equivalent. Therefore, we use Ψ as a shorthand to refer to any (arbitrarily chosen) among them. We can construct a set Ψ using a bottom-up traversal of the tree, as we have done elsewhere (Mai and Mirarab, 2020).

2.2 Maximum likelihood estimation using EM algorithm

Under the categorical model of the rates and the Gaussian model of branch length estimation error, the log-likelihood of $\hat{b_i}$ given the parameters τ and ω is

$$l(\tau,\omega) = \sum_{i=1}^{N} \log L_i(\hat{b_i};\tau_i,\omega) = \sum_{i=1}^{N} \log \sum_{j=1}^{k} f(\hat{b_i};\omega_j,\tau_i) Pr(\mu_i = \omega_j;\omega) = \sum_{i=1}^{N} \log \left(\frac{1}{k} \sum_{j=1}^{k} f(\hat{b_i};\omega_j,\tau_i) \right) \tag{3}$$

where $L_i(\hat{b_i};\tau_i,\omega)$ denotes the density of $\hat{b_i}$ on branch i and f is the density function of the Gaussian model. Our goal is to find τ and ω that maximize $l(\tau,\omega)$ and satisfy Ψ. Because l has a summation inside the log function, it is difficult to optimize it directly. However, thanks to the categorical model of the rates, the latent variables μ_i are discrete, and we can readily apply the EM algorithm (Dempster et al., 1977). We start with an initial of τ and ω (construction of these is described later) and iteratively improve the log-likelihood function by alternating between the E-step and M-step. In the **E-step**, we compute the posterior of the latent variables, setting

$$q_{ij} = Pr(\mu_i = \omega_j | \hat{b_i};\tau,\omega) = \frac{f(\hat{b_i}|\mu_i=\omega_j,\tau_i) Pr(\mu_i=\omega_j;\tau,\omega)}{\sum_{m=1}^{k} f(\hat{b_i}|\mu_i=\omega_m,\tau_i) Pr(\mu_i=\omega_m;\tau,\omega)} = \frac{f(\hat{b_i}|\mu_i=\omega_j,\tau_i)}{\sum_{m=1}^{k} f(\hat{b_i}|\mu_i=\omega_m,\tau_i)}$$

where the second equality holds because $Pr(\mu_i = \omega_m;\tau,\omega) = \frac{1}{k}$ for all i, m. Note these values are trivial to compute as f is simply the Gaussian pdf. In the **M-step**, we find ω and τ to maximize

$$\sum_{i=1}^{N} \sum_{j=1}^{k} q_{ij} \log f(\hat{b_i}|\omega_j,\tau_i), \tag{4}$$

s.t. the constraints Ψ given in (2) are satisfied. Using the Gaussian pdf and removing constants, the problem is to find:

$$\min_{\tau>0,\omega>0} \sum_{i=1}^{N} \sum_{j=1}^{k} \frac{sq_{ij}}{\hat{b_i}} (\hat{b_i} - \omega_j \tau_i)^2 \tag{5}$$

such that Ψ are satisfied. This optimization problem is non-convex and is difficult to solve exactly. However, it is easy to see that the likelihood function is bounded above, so it has a maximum. Therefore, the EM algorithm still converges as long as after every iteration (h) the M-step finds a new point $(\tau^{(h+1)}, \omega^{(h+1)})$ that gives a higher value for Eq. (4) than that of $(\tau^{(h)}, \omega^{(h)})$. Thus, it is sufficient to find a local minimum of Eq. (5) in every iteration to guarantee the convergence of the EM algorithm (a proof of convergence is shown in Supplementary). We use coordinate descent to find multiple local minima of Eq. (5) and select the one that gives the lowest penalty. In the M-step, starting with $\omega^{(h)}$ and $\tau^{(h)}$, we successively minimize Eq. (5) along the coordinate block of either ω or τ while fixing the other and iterate until convergence. In other words, let $\tau^{(h,1)} = \tau^{(h)}$ and $\omega^{(h,1)} = \omega^{(h)}$; in each iteration (p) of coordinate descent, we compute $\tau^{(h,p+1)}$ and $\omega^{(h,p+1)}$ such that:

$$\tau^{(h,p+1)} = \underset{\tau\geq 0, \tau \text{satisfies} \Psi}{\arg\min} \sum_{i=1}^{N} \sum_{j=1}^{k} \frac{sq_{ij}}{\hat{b_i}} (\hat{b_i} - \omega_j^{(h,p)} \tau_i)^2$$

$$\omega^{(h,p+1)} = \underset{\omega\geq\epsilon}{\arg\min} \sum_{i=1}^{N} \sum_{j=1}^{k} \frac{sq_{ij}}{\hat{b_i}} (\hat{b_i} - \omega_j \tau_i^{(h,p+1)})^2$$

These two problems are instances of the *weighted least-square* optimization that can be solved efficiently using convex programming. Inspired by To et al. (2016), we use the active-set method to solve these two problems. The complexity of each iteration of the active-set method is $O(N+k)$ (Supplementary Materials).

Initialization. The EM algorithm can only find local optima, so it needs multiple initial points. To facilitate the search, we first estimate the expected mutation rate μ, then discretize the uniform distribution $[0, 2\mu]$ into k equal segments and set $\omega_j^{(0)}$ to the middle of the jth segment. In other words,

$$\omega_j^{(0)} = (2j-1)\frac{\mu}{k}, \forall j \in [k] \ . \tag{6}$$

To initialize $\tau_i^{(0)}$, we draw j uniformly in $[k]$ and set $\tau_i^{(0)} = \frac{\hat{b}_i}{\omega_j^{(0)}}$. Although this initialization does not guarantee that $\tau_i^{(0)}$ satisfy Ψ, the constraints will be satisfied after the first M-step. We run wLogDate (Mai and Mirarab, 2020) and root-to-tip (RTT) regression (Shankarappa et al., 1999) to get two different estimates of μ. We then use each of these two estimates to get m different initials for $\tau^{(0)}$, for a total of $2m$ initials. We use $m = 100$ by default.

2-round optimization. A difficulty in co-estimating rates and times is that they are inseparable as a product. This same problem occurs in the EM algorithm; in each iteration, ω and τ can both be scaled up or down by the same factor, only constrained by calibration points. To avoid rapid jumps of ω and τ, we use a two-step approach (Fig. 1d). We first enforce ω to have the mean μ (estimated by either wLogDate or RTT, as described above) and run EM until convergence. Then we relax that constraint and let EM continue running to allow the expected μ to be re-estimated. Thus, for each initial point, we run EM twice. In the first round, the expected mutation rate is fixed to the initial μ by adding the constraint $\sum_j \omega_j = k\mu$ to Ψ. In the second round, we initialize EM with the ω and τ found in the first round and relax the constraint $\sum_j \omega_j = k\mu$ and let the EM algorithm run until convergence. The final solution is the one that gives the lowest value for Eq. (5) among all initial points. Consistent with the theory, the log-likelihood always increases monotonically after each iteration for all of the initial points (Fig. S4 in the Supplementary).

2.3 Evaluation: datasets

We used a simulated and a real angiosperm dataset and a simulated and a real HIV *env* gene dataset to evaluate MD-Cat in comparison to existing methods, LSD, BEAST, wLogDate, and RelTime.

Simulated Angiosperms hybrid rate

Flowering plants (angiosperms) is a well-studied example of the discrepancy between molecular dating and fossil records. Recent studies using molecular sequences report the estimate of the angiosperm crown age to be around 190-250 Mya (Clarke et al., 2011; Foster et al., 2017; Zeng et al., 2014), i.e., Triassic or earlier, while the oldest fossils date around 140 Mya, leaving a gap of as much as 100 million years between molecular dating and fossil records. Beaulieu et al. (2015) studied the effect of hybrid mutation rates on the accuracy of dating angiosperms. They simulated hybrid rate models for a phylogeny of seed plants where evolutionary rates formed local clocks in certain (groups of) clades. Their results showed that Bayesian estimation can perform poorly on such a complex clock model, providing a possible explanation for the gap between molecular dating and fossil records. In their simulations, the true age in all replicates was fixed to 140 mya. For rates, there are 5 different scenarios where the relative ratios of the rates were changed. See supplementary Table S2 and Beaulieu et al. (2015) for details of the 5 scenarios and note that going from scenario 1 to 5, rates become more varied. The time tree and 100 simulated phylograms for each of these five scenarios were downloaded from the Dryad Repository provided by the authors. We used the provided phylograms to simulate DNA sequences of length 1000 using SeqGen under GTR model with the shape of the Gamma rate heterogeneity across sites set to $\alpha = 1$. We compared MD-Cat to three other methods: BEAST (Drummond and Rambaut, 2007), wLogDate (Mai and Mirarab, 2020), and RelTime (Tamura et al., 2012). We used the true rooted tree topology and inferred branch lengths in substitution units using RAxML (Stamatakis, 2006). Because dating methods required a rooted tree as input, we used the 20 species on the clade outside the Angiosperm as outgroups to root the RAxML tree. Because outgroup rooting cannot determine the exact root position on the branch connecting ingroups and outgroups, we removed this clade from the tree before dating. As such, the 5 calibration points that belong to the 20 species in the outgroups were discarded, leaving us with 15 calibration points to be used in this analysis. We dated the tree using MD-Cat, wLogdate, and RelTime, providing each with the exact time of those 15 calibration points. We ran wLogDate with 100 feasible starting points, MD-Cat with $k = 50$ rate categories and 200 initial points, and RelTime with default settings. We ran BEAST with the topology fixed to the true topology, used the same settings as Beaulieu et al. (2015), and set the MCMC chain to 30 million. Our ESS for the posterior root age is well above 100 for a majority of the replicates (Fig. S2 in Supplementary), showing evidence of convergence.

Simulated HIV phylodynamics

We reused the phylodynamics data of HIV *env* gene simulated by To et al. (2016) but explored many more clock models. To et al. (2016) simulated time trees based on a birth-death model with periodic sampling times. There are four tree models: D995_11_10 (M1) and D995_3_25 (M2) simulate intra-host HIV evolution and D750_11_10 (M3) and D750_3_25 (M4) simulate inter-host evolution. Each tree model has 100 replicates, so we have 400 different tree topologies in total. Using these time trees, in our earlier work (Mai and Mirarab, 2020), we simulated substitution unit trees using three clock models: Lognormal, Gamma, and Exponential. In this paper, we further augmented this dataset by nine new clock models: one uniform distribution and eight mixtures of two, three, or four Lognormal distributions (Fig. 3a and Table S1 in Supplementary). We simulated sequence data using Seqgen under the same settings as the original study (To et al., 2016): sequence length is 1000; DNA evolution model is the F84 model with a gamma distribution for across-sites rate heterogeneity with shape 1.0 and eight rate categories; transition/transversion rate ratio is 2.5; nucleotide frequency of (A, C, G, T) is $(0.35, 0.20, 0.20, 0.25)$. We compared MD-Cat to LSD (To et al., 2016), wLogDate (Mai and Mirarab, 2020) and BEAST (Drummond and Suchard, 2010). We used the true rooted tree topology and only inferred branch lengths in time units. To test LSD, wLogDate, and MD-Cat, we used RAxML (Stamatakis, 2006) to estimate the branch lengths in substitution units from the simulated sequences (using the GTRGAMMA model) and used each of these methods to infer the time tree. LSD was run on the same settings as the QPD* mode described in the original paper (To et al., 2016). wLogDate was run with 100 feasible starting points. MD-Cat was run with $k = 50$ rate categories and 200 initials. To test BEAST, we used the sequences simulated by SeqGen, also fixed the true rooted tree topology and only inferred node ages. We ran BEAST using the following priors: HKY+Γ8 model, coalescent with constant population size, and strict-clock (i.e. fixed rate) clock model. We set the length of the MCMC chain to 5×10^6 generations, burn-in to 10%, and sampling to every 10^4 generations.

Real angiosperm data

We used MD-Cat to reanalyze the biological angiosperm data by Beaulieu et al. (2015) that consists of DNA sequences of four genes (atpB, rbcL, psbA, and 18S). We reused the rooted tree topology estimated by BEAST from the original study and ran RAxML on the concatenated alignment to estimate the branch lengths in substitution units (fixing the topology). We then used MD-Cat to estimate the divergence times and per-branch mutation rates. We used the same 24 calibration points as Beaulieu et al. (2015), except for the calibration for angiosperm, which is the node that we would like to evaluate. Each calibration point was incorporated as a hard constraint (i.e., equality).

HIV LANL dataset

We retrieved the alignment of the *env* region for HIV-1 M group, which includes subtypes A–L, from the Los Alamos HIV Sequence Database, LANL (http://www.hiv.lanl.gov/). The dataset includes 4062 sequences sampled from 1981 to 2020, with sampling year provided (328 sequences that have missing data were removed). We randomly subsampled subtypes B and C such that each of them has at most 20 sequences for each distinct year and kept all sequences of the other subtypes. As such, 1104 sequences remained, consisting of 2 subtype A, 208 subtype A1, 4 subtype A2, 447 subtype B, 200 subtype C, 99 subtype D, 41 subtype F1, 13 subtype F2, 79 subtype G, and 11 sequences of subtypes H-L. We used IQTree (with ModelFinder, which selected the GTR+F+I+I+R10 model) to estimate the tree topology and branch lengths, then rooted the tree using root-to-tip regression (Mai et al., 2017). Because 2 subtype A sequences did not form a monophyletic clade in our tree, we removed them together with one subtype A1 sequence that mixed with the subtype A2. After these steps, we got a rooted tree of 1101 sequences which has all the monophyletic subtypes. We used MD-Cat and the available sampling years to date the tree.

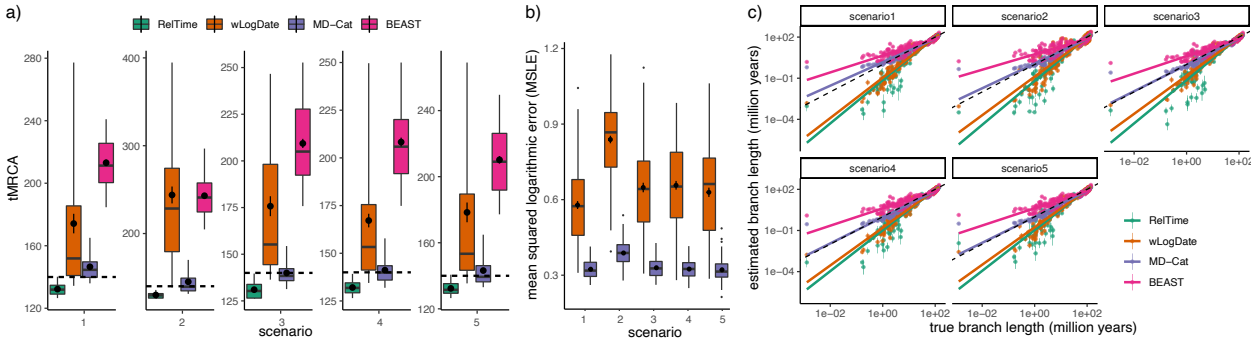

Figure 2: Comparison on the simulated Angiosperm dataset. (a) The tMRCA estimated by each method on each of the 5 scenarios. The dashed line shows the true tMRCA at 140. Boxplots, means (black dots), and standard error (bars) are shown over 100 replicates. (b) Mean-squared logarithmic error (MSLE) of the per-branch estimated mutation rates of wLogDate and MD-Cat (base 10) shown as boxplots, mean, and standard error over 100 replicates. (c) Estimated versus true branch lengths in time units. Mean and standard error are shown for the estimates across 100 replicates per branch. Both axes are shown in log-scale. Dashed line shows the unity line (i.e., perfect accuracy).

3 Results

3.1 Simulated angiosperms hybrid rate

In terms of tMRCAs estimation, MD-Cat outperforms wLogDate, RelTime, and BEAST in all five rate scenarios (Fig. 2a). Both BEAST and wLogDate severely overestimate the Angiosperm age for all scenarios. RelTime has a lower error than BEAST and wLogDate, but it underestimates the age. In contrast, MD-Cat has both low error (5.2% in average) and low bias: for scenario 1, it overestimates the age by 5 million years; for scenarios 2–5, it underestimates by 0.5, 1.6, 0.1, and 0.6 million years, respectively. Overall, MD-Cat is more accurate and less biased than the other methods. In terms of the per-branch mutation rates, MD-Cat dramatically outperforms wLogDate, the only other method that provides such rates (Fig. 2b).

Comparison of estimated branch lengths in time unit confirms the accuracy of MD-Cat (Fig. 2c). wLogDate has a systematic bias: it underestimates short branches but tends to overestimate the longer branches. This overestimation of some long branches explains the overestimation of the tMRCA, despite the fact that the method underestimates a majority of the branches. RelTime has an overall tendency to underestimate the branch lengths, which is consistent with its underestimation of the tMRCA. Moreover, the estimates by RelTime have high variance, where branches of similar true lengths can be assigned very different estimated lengths. At the other extreme, BEAST consistently overestimates the branch lengths for all scenarios. Overall, the estimates by MD-Cat are closer to the true values and have both lower variance and bias.

3.2 Simulated HIV phylodynamics

All methods estimate the tMRCA most accurately on the two simple unimodal models (Gamma and LogNormal), and substantially less accurately on the non-modal (Exponential and uniform) and multimodal models (Fig. 3). The long-tailed exponential distribution and all the multimodal distributions except for Biomodal 1 and Trimodal 1 seem particularly challenging (Fig. 2b). The two multimodal distributions that are less challenging are symmetric around their mean and have less distance between modes. Distributions with long tails (compared to the mean) seem particularly challenging for all the methods.

Compared to the other methods, MD-Cat estimates tMRCA with comparable or better accuracy across all conditions (Fig. 3c). As expected, the strict-clock method, LSD, performs well on the unimodal cases (Gamma and Lognormal), but its accuracy drops dramatically on the other clock models. While wLogDate and BEAST are more robust than LSD, they both overestimate the tMRCA across all clock models (the average bias is 5.0% and 8.9% for wLogDate and BEAST, respectively, averaged over all models) and especially for the nonmodal and multi-modal distributions. For example, while the biases are less than 2.5% for both methods under the LogNormal distribution, under the challenging Binomial 4 model, the mean bias is 15.5% and 9.8% for BEAST and wLogDate, respectively. MD-Cat overestimates the tMRCA in 7 models, among which only Exponential has a somewhat high bias (6.7%), and the rest (Gamma, LogNormal, Uniform,

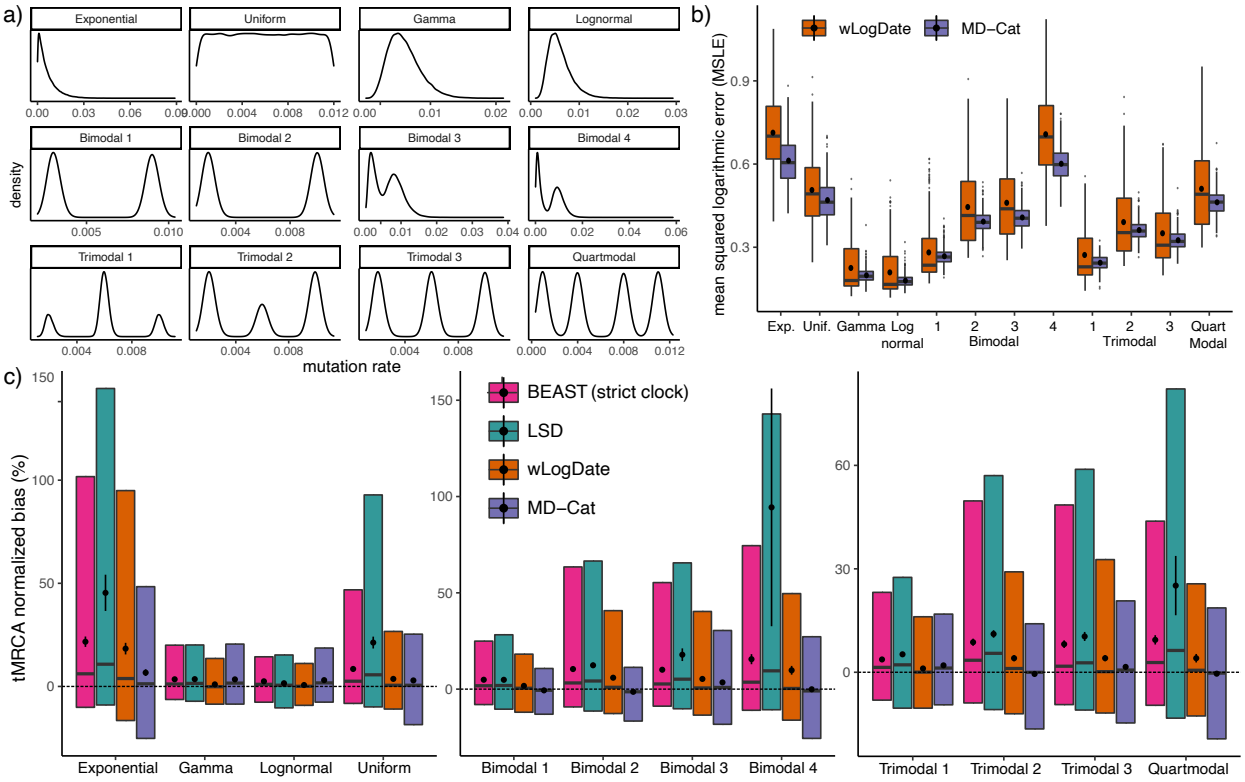

Figure 3: Comparison on the HIV simulated dataset. (a) The empirical distribution of the mutation rates simulated for each of the 12 clock models. (b) Mean-squared logarithmic (base 10) error (MSLE) of the per-branch estimated mutation rates of wLogDate and MD-Cat shown as boxplots, mean, and standard error over 400 replicates. (c) The bias in estimated root divergence time (tMRCA). The true tMRCA is fixed at 0. Model conditions are divided into those with zero or one mode (left), two modes (middle), and three or more modes (right). Boxes show median and 5-95 percentile instead of standard boxplot settings due to high variance; mean and standard error are also shown.

Bimodal 3, Trimodal 1, and Trimodal 3) have a bias below 3.5%. In the other 5 conditions, MD-Cat slightly underestimates tMRCA, with the average bias at 1.4% for Bimodal 2 and below 1% in other cases (Bimodals 1 and 4, Trimodal 2, and Quartmodal). Averaging across all models, MD-Cat has a small overestimation bias of 1.7%. MD-Cat also has fewer cases of extreme error; e.g., tMRCA error of more than 50% happens only 50 times across all conditions with MD-Cat but 213 and 137 times with BEAST and wLogDate. The error of divergence time estimation (i.e. root-mean-squared-error divided by the tree height) shows similar patterns (Fig. S4 in Supplementary). In the two unimodal models, all methods are relatively accurate and have similar accuracy. MD-Cat is more accurate than the other methods in either non-modal or multi-modal in the other clock models. The average improvements of MD-Cat over the second-best method range from small (e.g., 0.3% for Trimodal 1 model) to substantial (3% or more for Exponential, Bimodal 2, and Bimodal 4). Across all clock models, the average errors of LSD, wLogDate, BEAST, and MD-Cat are 15%, 8.0%, 8.1%, and 6.7%, respectively. For example, the 95-percentile error of LSD, BEAST, and wLogDate for the Exponential model are 70%, 48%, and 50%, resp., compared to 22% for MD-Cat. Comparing the estimated mutation rates by wLogDate and MD-Cat confirms the same relative performance of the two methods, where MD-Cat has a lower average error in all clock models (Fig. 3b).

Comparing the running times, LSD is extremely fast, finishing each replicate in less than a second. Among the other methods (Fig. S1 in Supplementary), wLogDate is the fastest method with an average running time of 15 minutes, followed by MD-Cat at 30 minutes, which is more than 5 times faster than BEAST at 159 minutes. There is no substantial difference in running time among different clock models.

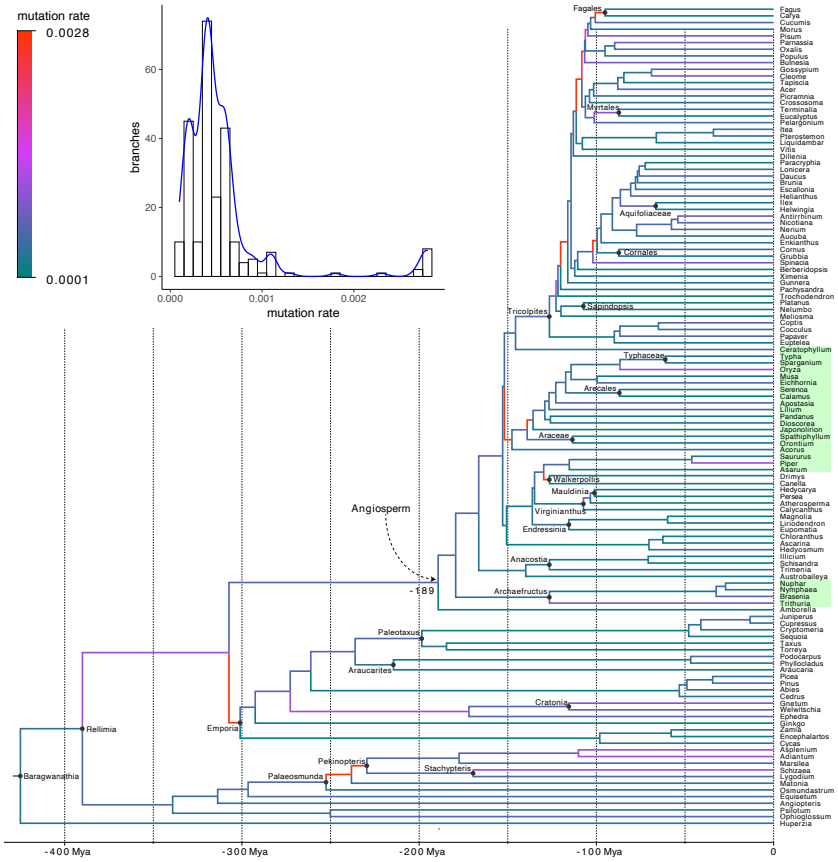

Figure 4: Dating land plants using MD-Cat. Branches are colored by the corresponding estimated mutation rate, with the histogram and density plot of the rates shown as inset. Calibration points are shown with solid dots and the corresponding name. The age of angiosperm is estimated at 189 Mya. Four herbaceous groups singled out by Beaulieu et al. (2015) (Nymphaeales, Piperales, Monocotyledonae, and Ceratophyllum) are highlighted in green.

3.3 Dating of the real angiosperm data

Beaulieu et al. (2015) showed that BEAST with the LogNormal clock model estimates a very old age for the crown angiosperm (median 232 Mya; 95% HPD 210–256), and this estimate is much older than evidence from fossils. In particular, the earliest available fossils reliably identified as ancestral angiosperms are dated to 140 mya. The 100 million years gap between the BEAST estimate and the earliest known fossil led to the suspicion that BEAST could be overestimating dates due to the use of an incorrect clock model. The authors concluded that sources of bias in BEAST mean that we cannot rule out a much more recent age, closer to 140 mya, than what BEAST estimates. MD-Cat estimates the divergence time of angiosperms to be 189 Mya (Fig. 4). Although this estimated age is still substantially older than the earliest available fossil, the gap is reduced by half relative to BEAST. Moreover, our estimate is within the range of previous estimates by molecular dating, which ranged between 175 and 251 Mya (Beaulieu et al., 2015; Clarke et al., 2011; Foster et al., 2017; Zeng et al., 2014), and is on the younger side. The estimated branch rates form a bimodal (or perhaps trimodal) distribution (Fig. 4) that would not fit commonly-used unimodal distributions such as LogNormal. This distribution has a substantial number of branches with much higher rates (close to 0.0025) than typical branches, which have rates distributed around 0.00055. Some of these high-rate branches are close to the root of angiosperms; e.g., the branch sister to the root of Angiosperms has a high rate. Among herbaceous angiosperms singled out by Beaulieu et al. (2015) as potential sources of increased rate (Nymphaeales, Piperales, Monocotyledonae, and Ceratophyllum), some but not all branches have clearly elevated rates; e.g., Trithuria at the base of Nymphaeales, piper among Piperales, and the domesticated Oryza among monocots have elevated rates. Most strikingly, the MRCA of all monocots and the MRCA of all non-Araceae monocots both have rates close to 0.0025, which is 5× the typical rate across

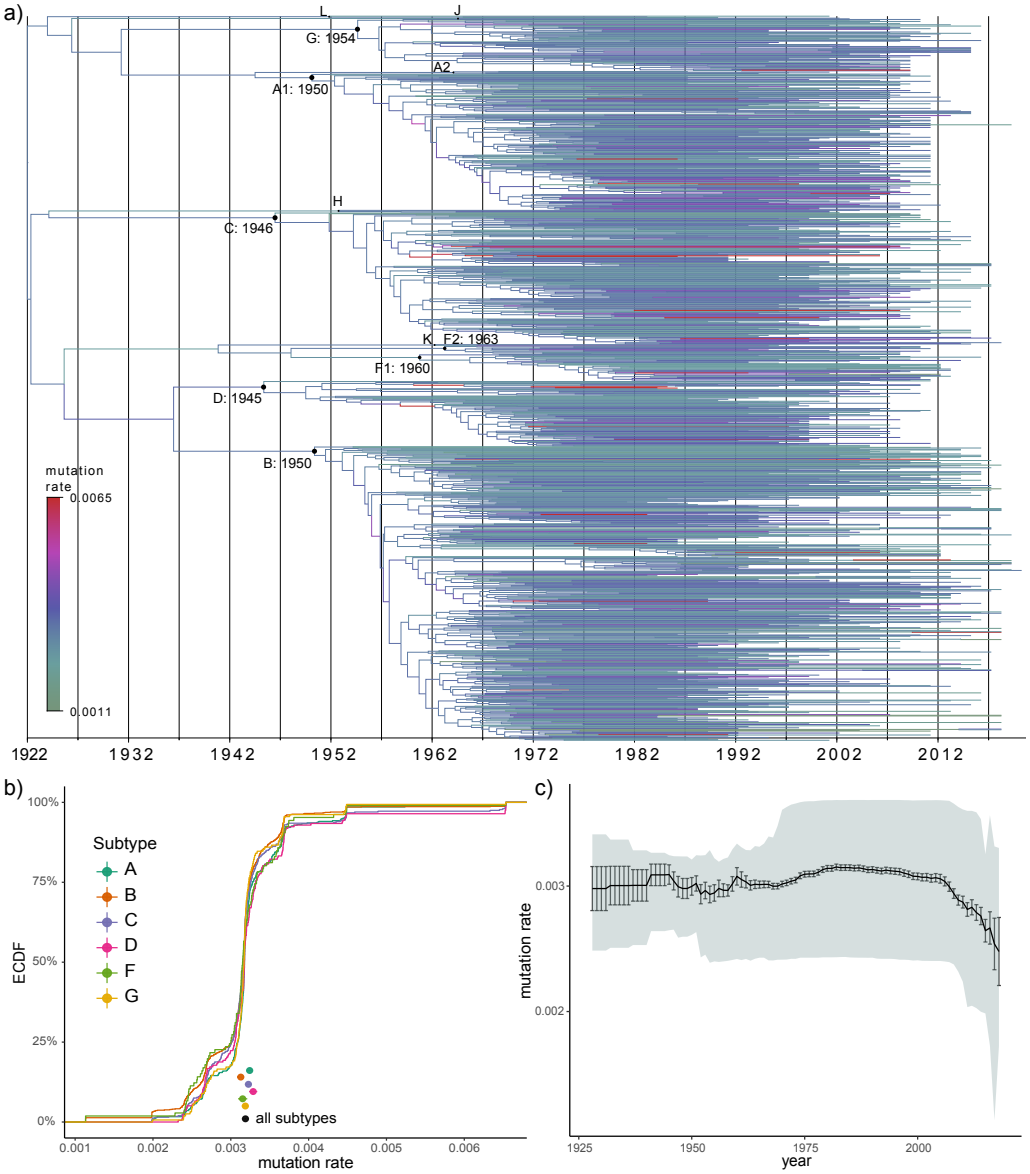

Figure 5: Dated HIV tree. a) The chronogram estimated by MD-Cat. Branches are colored by the estimated substitution rate. b) The distribution of rates separated for the major subtypes, shown as an empirical cumulative distribution (ECDF). The mean and standard error of each subtype and all subtypes combined are also shown as dots and error bars (with arbitrary positions along the y-axis); only branches below the MRCA of each subtype are included. c) Rates across time. For each year, we computed the distribution of rates of all branches that span that entire year; we show the mean and standard error as a line and error bars and 10–90% quantiles as shaded area.

the tree. These branches, close to the base of angiosperms, likely have a major impact on its estimated tMRCA; lowering their rate would put the angiosperm tMRCA further back in time.

3.4 Dating of the real combined HIV subtypes

Our estimate of the tMRCA of the entire HIV-1 M group is 1922, consistent with the hypothesis about the Kinshasa spread of HIV-1 in 1920s proposed by many analyses (Faria et al., 2014; Junqueira and de Matos Almeida, 2016; Keele et al., 2006; Korber et al., 2000). Also, according to these earlier studies, the subtype B lineage originated in Kinshasa during 1940s but was detected outside Africa only in 1960s (Faria et al., 2014; Junqueira and de Matos Almeida, 2016). Molecular dating methods often place the tMRCA for subtype B in between these two events, i.e., in 1950s (Abecasis et al., 2009; Korber et al.,

2000; Patiño-Galindo and González-Candelas, 2017; Wertheim et al., 2012). Consistent with these results, MD-Cat estimates the tMRCA of subtype B to be 1950, which is close to the proposed true origin in 1940s. In addition, our estimates of the tMRCA of most of the other subtypes are also within the confidence interval of previous works (Fig. 5a and Table S3). The average mutation rate across all subtypes is estimated at 0.0032, and we observed no substantial difference among the subtypes (Fig. 5b). Interestingly, tracking the mutation rates through time (Fig. 5c) by averaging all branches spanning each year showed consistent estimates across years until the early 2000s when the rate started to drop. This period coincides with the wider availability of HIV antiviral drugs, which can stop the evolution of the virus while taken by an individual. Thus, the treatments may lead to lowered observed mutation rates at the population level (evolution rebounds after/when individuals stop taking the treatment, which happens often). However, this trend can also be an artifact of the lower number of branches spanning recent years (and hence the higher variance in recent years).

4 Discussions

We introduced a phylogenetic dating method using the flexible categorical model. Although CAT models are not new in phylogenetics, their power to approximate a continuous distribution as the clock model has not been studied before. We formulate the dating problem as maximum likelihood inference and solve it using an EM-based algorithm. Although the likelihood function is non-convex, the objective function in the M-step is convex on ω and τ separately. We use block coordinate descent to alternatively optimize ω and τ, and show that each iteration can be solved efficiently. Together, these strategies greatly reduce the running time of the MD-Cat method. On two simulated datasets of the Angiosperms and HIV, MD-Cat outperforms other methods in most clock models, especially in the model conditions that violate their (implicit) assumptions of a strict or unimodal clock model. Improvements are most visible under the hardest model conditions where either the sudden rate shift is extreme, or the overall rate distribution has high variance.

Despite its improved empirical accuracy, many theoretical and computational questions related to MD-Cat remain unanswered. Foremost, MD-Cat can suffer from over-parameterization. As noted before, time and rate are inseparable from molecular data, even with the aiding information from calibration points. In general, approximating a continuous distribution requires a large number of rate categories (k). However, the data size is small compared to the number of unknown parameters ($2n-2$ branch length observations and $O(n)$ calibration points versus $2n-2+k$ parameters). Therefore, the parameters are unidentifiable in general, pointing to the difficulty of getting the correct age for every node. With the pre-estimate of the expected mutation rate and our 2-round optimization strategy, we have tried to address the over-parameterization in MD-Cat. Nevertheless, our evidence that the inference of a substantial number of rates is doable despite the large space is mostly empirical. We leave the theoretical analysis of identifiability of ω and τ under this categorical model and different combinations of calibration points for future work. While we formulate and estimate ω and τ under a maximum likelihood framework, we note that the proposed categorical model can also be used in Bayesian inference, where a prior for the rate categories can be specified. For scalability purposes, we use the inferred tree in substitution units as input and use a simple Gaussian model for branch length estimation error. A more direct method is to incorporate the categorical clock model into the tree inference from sequence data and maximize the joint likelihood function. While this approach avoids assuming a Gaussian model for branch length estimation error, we note that it may exacerbate the over-parameterization problem, as all the parameters (tree topology, GTR parameters, per-site, and per-branch rate heterogeneity) must be co-estimated. Future works should explore this approach, both theoretically and empirically. In addition, rooting phylogenies is a non-trivial problem and is also related to the problem of clock model selection. The current formulation of MD-Cat only works on rooted phylogenetic trees, but the generalization of MD-Cat to unrooted trees is a straight-forward: solve the optimization problem for each possible rooting and select the root position that has the maximum likelihood. Such an approach should be explored in future works, together with an updated formulation for MD-Cat to add a parameter that determines the optimal root placement on each branch.

Our study can also be extended in the future. To facilitate the comparison between different methods, we used the true topology with estimated branch lengths and left the study of the impact of incorrect topology on the dating methods for future works. Finally, testing under more complex clock models, such as those that

allow continuous rate change, is also a topic for future works. Our analyses of real data provided interesting insights, including the idea that mutation rates of HIV may have dropped. Further studies should examine if the reduced rate-through-time trend is real and if it is connected to any epidemiological factors.

References

A. B. Abecasis, A.-M. Vandamme, and P. Lemey. Quantifying differences in the tempo of human immunodeficiency virus type 1 subtype evolution. *Journal of virology*, 83(24):12917–12924, 2009.

S. Aris-Brosou and Z. Yang. Effects of models of rate evolution on estimation of divergence dates with special reference to the metazoan 18s ribosomal RNA phylogeny. *Systematic Biology*, 51(5):703–714, Sept. 2002. doi: 10.1080/10635150290102375. URL https://doi.org/10.1080/10635150290102375.

J. M. Beaulieu, B. C. O'Meara, P. Crane, and M. J. Donoghue. Heterogeneous Rates of Molecular Evolution and Diversification Could Explain the Triassic Age Estimate for Angiosperms. *Systematic Biology*, 64 (5):869–878, 05 2015. ISSN 1063-5157. doi: 10.1093/sysbio/syv027. URL https://doi.org/10.1093/sysbio/syv027.

T. Britton, C. L. Anderson, D. Jacquet, S. Lundqvist, and K. Bremer. Estimating Divergence Times in Large Phylogenetic Trees. *Systematic Biology*, 56(5):741–752, 10 2007. ISSN 1063-5157. doi: 10.1080/10635150701613783. URL https://doi.org/10.1080/10635150701613783.

L. Bromham and D. Penny. The modern molecular clock. *Nature Reviews Genetics*, 4(3):216–224, 3 2003. ISSN 1471-0056. doi: 10.1038/nrg1020. URL http://www.nature.com/articles/nrg1020.

J. T. Clarke, R. C. Warnock, and P. C. Donoghue. Establishing a time-scale for plant evolution. *New Phytologist*, 192(1):266–301, 2011.

A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, 39(1):1–38, 1977. ISSN 00359246. URL http://www.jstor.org/stable/2984875.

A. J. Drummond and A. Rambaut. BEAST: Bayesian evolutionary analysis by sampling trees. *BMC evolutionary biology*, 7:214, 2007. ISSN 1471-2148. doi: 10.1186/1471-2148-7-214.

A. J. Drummond and M. A. Suchard. Bayesian random local clocks, or one rate to rule them all. *BMC Biology*, 8(1):114, 12 2010. ISSN 1741-7007. doi: 10.1186/1741-7007-8-114. URL https://bmcbiol.biomedcentral.com/articles/10.1186/1741-7007-8-114.

A. J. Drummond, S. Y. W. Ho, M. J. Phillips, and A. Rambaut. Relaxed Phylogenetics and Dating with Confidence. *PLoS Biology*, 4(5):e88, 3 2006. ISSN 1545-7885. doi: 10.1371/journal.pbio.0040088. URL https://dx.plos.org/10.1371/journal.pbio.0040088.

N. R. Faria, A. Rambaut, M. A. Suchard, G. Baele, T. Bedford, M. J. Ward, A. J. Tatem, J. D. Sousa, N. Arinaminpathy, J. Pépin, et al. The early spread and epidemic ignition of hiv-1 in human populations. *science*, 346(6205):56–61, 2014.

J. Felsenstein. Phylip (phylogeny inference package) version 3.6. distributed by the author. *http://www. evolution. gs. washington. edu/phylip. html*, 2004.

C. S. Foster, H. Sauquet, M. Van der Merwe, H. McPherson, M. Rossetto, and S. Y. Ho. Evaluating the impact of genomic data and priors on bayesian estimates of the angiosperm evolutionary timescale. *Systematic Biology*, 66(3):338–351, 2017.

S. Guindon. Bayesian Estimation of Divergence Times from Large Sequence Alignments. *Molecular Biology and Evolution*, 27(8):1768–1781, 03 2010. ISSN 0737-4038. doi: 10.1093/molbev/msq060. URL https://doi.org/10.1093/molbev/msq060.

S. Guindon, J.-F. Dufayard, V. Lefort, M. Anisimova, W. Hordijk, and O. Gascuel. New Algorithms and Methods to Estimate Maximum-Likelihood Phylogenies: Assessing the Performance of PhyML 3.0. *Systematic Biology*, 59(3):307–321, 5 2010. doi: 10.1093/sysbio/syq010. URL http://sysbio.oxfordjournals.org/content/59/3/307.abstracthttp://sysbio.oxfordjournals.org/content/59/3/307.full.pdfhttp://sysbio.oxfordjournals.org/cgi/content/abstract/59/3/307.

T. A. Heath, J. P. Huelsenbeck, and T. Stadler. The fossilized birth–death process for coherent calibration of divergence-time estimates. *Proceedings of the National Academy of Sciences*, 111(29):E2957–E2966, 2014.

S. Y. Ho and S. Duchêne. Molecular-clock methods for estimating evolutionary rates and timescales. *Molecular ecology*, 23(24):5947–5965, 2014.

D. M. Junqueira and S. E. de Matos Almeida. Hiv-1 subtype b: Traces of a pandemic. *Virology*, 495:173–184, 2016.

B. F. Keele, F. Van Heuverswyn, Y. Li, E. Bailes, J. Takehisa, M. L. Santiago, F. Bibollet-Ruche, Y. Chen, L. V. Wain, F. Liegeois, et al. Chimpanzee reservoirs of pandemic and nonpandemic hiv-1. *Science*, 313

(5786):523–526, 2006.

B. Korber, M. Muldoon, J. Theiler, F. Gao, R. Gupta, A. Lapedes, B. Hahn, S. Wolinsky, and T. Bhattacharya. Timing the ancestor of the hiv-1 pandemic strains. *science*, 288(5472):1789–1796, 2000.

S. Kumar. Molecular clocks: four decades of evolution. *Nature Reviews Genetics*, 6(8):654–662, 8 2005. ISSN 1471-0056. doi: 10.1038/nrg1659. URL http://www.nature.com/articles/nrg1659.

S. Kumar and S. B. Hedges. Advances in Time Estimation Methods for Molecular Data. *Molecular Biology and Evolution*, 33(4):863–869, 02 2016. ISSN 0737-4038. doi: 10.1093/molbev/msw026. URL https://doi.org/10.1093/molbev/msw026.

C. H. Langley and W. M. Fitch. An examination of the constancy of the rate of molecular evolution. *Journal of Molecular Evolution*, 3(3):161–177, Sep 1974. ISSN 1432-1432. doi: 10.1007/BF01797451. URL https://doi.org/10.1007/BF01797451.

V. Lefort, R. Desper, and O. Gascuel. FastME 2.0: A Comprehensive, Accurate, and Fast Distance-Based Phylogeny Inference Program. *Molecular Biology and Evolution*, 32(10):2798–2800, 06 2015. ISSN 0737-4038. doi: 10.1093/molbev/msv150. URL https://doi.org/10.1093/molbev/msv150.

U. Mai and S. Mirarab. Log Transformation Improves Dating of Phylogenies. *Molecular Biology and Evolution*, 38(3):1151–1167, 09 2020. ISSN 1537-1719. doi: 10.1093/molbev/msaa222. URL https://doi.org/10.1093/molbev/msaa222.

U. Mai, E. Sayyari, and S. Mirarab. Minimum variance rooting of phylogenetic trees and implications for species tree reconstruction. *PLOS ONE*, 12(8):e0182238, 8 2017. ISSN 1932-6203. doi: 10.1371/journal.pone.0182238. URL http://dx.plos.org/10.1371/journal.pone.0182238.

L. T. Nguyen, H. A. Schmidt, A. Von Haeseler, and B. Q. Minh. IQ-TREE: A fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Molecular Biology and Evolution*, 32(1), 2015. ISSN 15371719. doi: 10.1093/molbev/msu300.

J. Á. Patiño-Galindo and F. González-Candelas. The substitution rate of hiv-1 subtypes: a genomic approach. *Virus Evolution*, 3(2), 2017.

B. Rannala and Z. Yang. Inferring Speciation Times under an Episodic Molecular Clock. *Systematic Biology*, 56(3):453–466, 06 2007. ISSN 1063-5157. doi: 10.1080/10635150701420643. URL https://doi.org/10.1080/10635150701420643.

F. Rutschmann. Molecular dating of phylogenetic trees: A brief review of current methods that estimate divergence times. *Diversity & Distributions*, 12(1):35–48, 1 2006. ISSN 1366-9516. doi: 10.1111/j.1366-9516.2006.00210.x. URL http://doi.wiley.com/10.1111/j.1366-9516.2006.00210.x.

P. Sagulenko, V. Puller, and R. A. Neher. TreeTime: Maximum-likelihood phylodynamic analysis. *Virus Evolution*, 4(1), 01 2018. ISSN 2057-1577. doi: 10.1093/ve/vex042. URL https://doi.org/10.1093/ve/vex042. vex042.

N. Saitou and M. Nei. The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Molecular Biology and Evolution*, 4(4):406–425, 07 1987. ISSN 0737-4038. doi: 10.1093/oxfordjournals.molbev.a040454. URL https://doi.org/10.1093/oxfordjournals.molbev.a040454.

M. J. Sanderson. A Nonparametric Approach to Estimating Divergence Times in the Absence of Rate Constancy. *Molecular Biology and Evolution*, 14(12):1218–1231, 12 1997. ISSN 0737-4038. doi: 10.1093/oxfordjournals.molbev.a025731. URL https://academic.oup.com/mbe/article-lookup/doi/10.1093/oxfordjournals.molbev.a025731.

M. J. Sanderson. Estimating rate and time in molecular phylogenies: beyond the molecular clock? In *Molecular systematics of plants II*, pages 242–264. Springer, 1998.

M. J. Sanderson. Estimating Absolute Rates of Molecular Evolution and Divergence Times: A Penalized Likelihood Approach. *Molecular Biology and Evolution*, 19(1):101–109, 1 2002. ISSN 1537-1719. doi: 10.1093/oxfordjournals.molbev.a003974. URL http://academic.oup.com/mbe/article/19/1/101/1066733.

R. Shankarappa, J. B. Margolick, S. J. Gange, A. G. Rodrigo, D. Upchurch, H. Farzadegan, P. Gupta, C. R. Rinaldo, G. H. Learn, X. He, X. L. Huang, and J. I. Mullins. Consistent viral evolutionary changes associated with the progression of human immunodeficiency virus type 1 infection. *Journal of virology*, 73(12):10489–502, 12 1999. ISSN 0022-538X. URL http://www.ncbi.nlm.nih.gov/pubmed/10559367http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC113104.

A. Stamatakis. RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. *Bioinformatics*, 22(21):2688–2690, 08 2006. ISSN 1367-4803. doi: 10.1093/bioinformatics/btl446. URL https://doi.org/10.1093/bioinformatics/btl446.

K. Tamura, F. U. Battistuzzi, P. Billing-Ross, O. Murillo, A. Filipski, and S. Kumar. Estimating divergence times in large molecular phylogenies. *Proceedings of the National Academy of Sciences*, 109(47):19333–19338, 11 2012. ISSN 0027-8424. doi: 10.1073/pnas.1213199109. URL http://www.pnas.org/cgi/doi/10.1073/pnas.1213199109.

K. Tamura, Q. Tao, and S. Kumar. Theoretical Foundation of the RelTime Method for Estimating Divergence Times from Variable Evolutionary Rates. *Molecular Biology and Evolution*, 35(7):1770–1782, 03 2018. ISSN 0737-4038. doi: 10.1093/molbev/msy044. URL https://doi.org/10.1093/molbev/msy044.

S. Tavaré. Some Probabilistic and Statistical Problems in the Analysis of DNA Sequences. *Lectures on Mathematics in the Life Sciences*, 17:57–86, 1986.

J. L. Thorne and H. Kishino. Divergence time and evolutionary rate estimation with multilocus data. *Systematic Biology*, 51(5):689–702, Sept. 2002. doi: 10.1080/10635150290102456. URL https://doi.org/10.1080/10635150290102456.

J. L. Thorne, H. Kishino, and I. S. Painter. Estimating the rate of evolution of the rate of molecular evolution. *Molecular Biology and Evolution*, 15(12):1647–1657, 12 1998. ISSN 0737-4038. doi: 10.1093/oxfordjournals.molbev.a025892. URL https://academic.oup.com/mbe/article-lookup/doi/10.1093/oxfordjournals.molbev.a025892.

T.-H. To, M. Jung, S. Lycett, and O. Gascuel. Fast Dating Using Least-Squares Criteria and Algorithms. *Systematic Biology*, 65(1):82–97, 1 2016. ISSN 1063-5157. doi: 10.1093/sysbio/syv068. URL https://academic.oup.com/sysbio/article-lookup/doi/10.1093/sysbio/syv068.

E. M. Volz and S. D. W. Frost. Scalable relaxed clock phylogenetic dating. *Virus Evolution*, 3(2), 08 2017. ISSN 2057-1577. doi: 10.1093/ve/vex025. URL https://doi.org/10.1093/ve/vex025. vex025.

J. O. Wertheim, M. J. Sanderson, M. Worobey, and A. Bjork. Relaxed Molecular Clocks, the Bias–Variance Trade-off, and the Quality of Phylogenetic Inference. *Systematic Biology*, 59(1):1–8, 1 2010. ISSN 1076-836X. doi: 10.1093/sysbio/syp072. URL https://academic.oup.com/sysbio/article/59/1/1/1723234.

J. O. Wertheim, M. Fourment, and S. L. Kosakovsky Pond. Inconsistencies in Estimating the Age of HIV-1 Subtypes Due to Heterotachy. *Molecular Biology and Evolution*, 29(2):451–456, 2 2012. ISSN 0737-4038. doi: 10.1093/molbev/msr266. URL https://academic.oup.com/mbe/article-lookup/doi/10.1093/molbev/msr266.

X. Xia and Q. Yang. A distance-based least-square method for dating speciation events. *Molecular Phylogenetics and Evolution*, 59(2):342 – 353, 2011. ISSN 1055-7903. doi: https://doi.org/10.1016/j.ympev.2011.01.017. URL http://www.sciencedirect.com/science/article/pii/S1055790311000443.

L. Zeng, Q. Zhang, R. Sun, H. Kong, N. Zhang, and H. Ma. Resolution of deep angiosperm phylogeny using conserved nuclear genes and estimates of early divergence times. *Nature communications*, 5(1):1–12, 2014.

E. Zuckerkandl. Molecular disease, evolution, and genetic heterogeneity. *Horizons in biochemistry*, pages 189–225, 1962.