Pending

John H Tay[1], Sebastian Duchene[1,2*].

[1]Peter Doherty Institute for Infection and Immunity, Department of Microbiology and Immunology, University of Melbourne, Melbourne, Australia.

[2]Department of Computational Biology, Institut Pasteur, Paris, France.

*email: sduchene@unimelb.edu.au

Abstract

Pending.

Outline

- Talk about tests of temporal signal
- The role of BETS
- pitfals of BETS: that we don't know the effect of the tree prior, all priors should be proper, impact of the clock model (because it can absorb rate variation among lineages).
- introduce data sets
- analyses of empirical data

  Use constant coalescent with the three population size priors and explain that they result in different priors on the clock rate and on the tree height. What are the implications here?

  Use the skyride and show the marginal prior on clock rate and tree height.
- Simulations. We should only do these for the best fitting model clock model per data set. But here, let's stick to the constant-size coalescent because it is more practical.

  How well do they classify data sets as having temporal signal? think about error rates.

  Does the classification accuracy depend on the average logBFs for those data sets that do have temporal signal? What I mean here is that the cholera ones would naturally have stronger temporal signal than the treponema ones, so we expect different error rates.
- Conclude: What is the best strategy here regarding tree prior? comment on the clock prior? where doe the regression fit in here? what about prior and posterior predictive checks?

- What does the root height, tree length, and clock rate prior look in each case (coalescent with all priors, skyride)?

  How do these priors impact the phylogenetic likelihood? The point here is that if the tree prior is very wrong (i.e. it is in conflict with the likelihood) then it can mislead tests of temporal signal.

  with the relaxed clock it is useful to inspect whether the coefficient of rate variation is larger

- Ideally the test should assess only the sampling times

Key points for Results section - 28 / 02 / 2024

- The impact of the tree prior and hyperpriors on model selection. Here show the polygons and tables. Pow is particularly interesting because the exponential prior on Φ does not support temporal signal. **Result: Priors obviously impact marginal likelihoods, but sometimes also model choice**

- Show clock rate (ucld.mean and clock.rate) and root height distributions under each clock model and prior. **Result: Estimates of evolutionary rates and times are generally robust to the tree prior even if the marginal likelihood is not**

- Simulations Pending for now **- does the tree prior impact the accuracy to detect temporal signal?**

- Why does the prior on Φ impact temporal signal? Because it is linked to tree length and to the clock rate via the coalescent and the CTMC reference prior. Here show pairs plots for the root height, Φ, tree length, clock rate **Result: Some parameters in the Bayesian hierarchical model are naturally correlated, the marginal prior is not obvious, but can be easily inspected**

- Inspecting the marginal prior is important and should be conducted prior to tests of temporal signal. For example, including monophyletic constraints, and other prior information can impact temporal signal.

- Prior sensitivity is also important to assess the robustness of the estimates.

Prior sensitivity to run (21 Feb)

- Run all three priors on Φ and check

  tree height, clock rate, tree length, pop size - and correlations

- Run with a set of fixed Φ values and show how the mean and variance of clock rate, tree height and tree length changes

new results, as of 12 Feb

- The prior on the population size seems to matter when we don't have temporal signal. It tends to lead to false positives.
- A data set with temporal signal, will generally be robust to the prior.
- Redefine temporal signal. It is when the data are more informative than the prior in a Bayesian context.
- But this is not a problem in itself. One should choose a prior that is sensible. Some data sets, like cholera will be robust to the choice of prior. They are more informative and have stronger temporal signal.

Introduction

Pending

Table 1: Proportion of simulations with temporal signal according to log Bayes factors support (log BF)

Prior on Φ and clock model	$logBF \geq 0$	$logBF \geq 1$	$logBF \geq 3$	$logBF \geq 5$
$\Gamma(\kappa = 0.001, \theta = 1000)$; SC	0.25	0.23	0.21	0.21
$Lnorm(\mu = 1.0, \sigma = 5)$; SC	0.42	0.40	0.33	0.26
$Exp(\mu = 1.0)$; SC	NA	NA	NA	NA

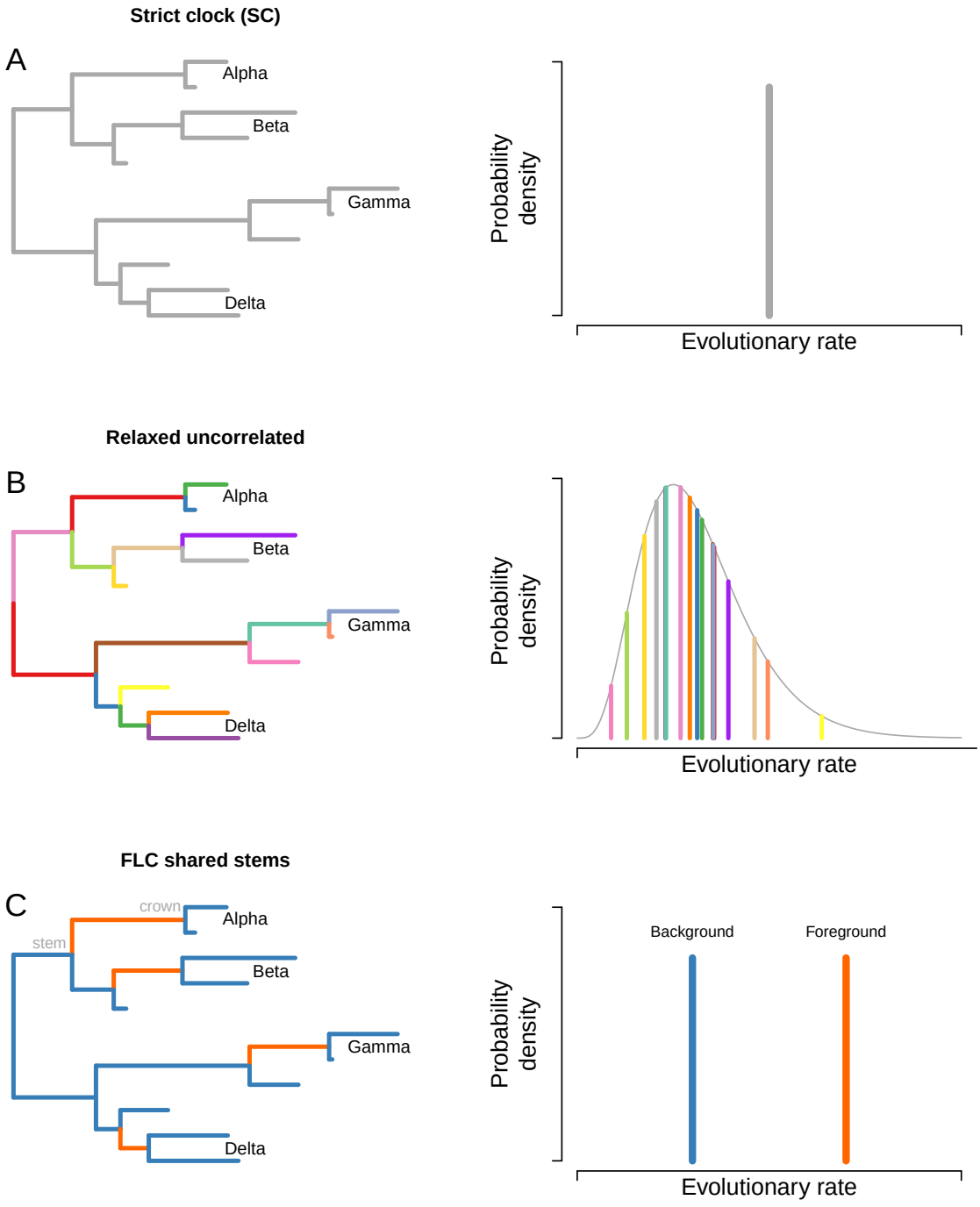

Figure 1: Examples of molecular clock models. A. is a strict clock model (SC), where all branches share a single evolutionary rate. B. is an uncorrelated relaxed clock model, in which the evolutionary rates across branches are independent draws from an underlying statistical distribution, such as a lognormal or a gamma (Γ) distribution (typically abbreviated as UCLD and UCGD, respectively). Note that branches are coloured according to those drawn from the distribution on the right. C. represents a fixed local clock model, where stem branches leading up to variants of concern (VOCs; those leading to clades labelled as Alpha, Beta, Gamma and Delta, following variant names) are designated as the 'foreground' and assigned a rate that differs from the 'background'. In C, we have labelled the stem and crown nodes of variant Alpha, where evolutionary rate changes would have occurred.

1 Supplementary Material

Supplementary data are available at Molecular Biology and Evolution online.

2 Acknowledgements

The Authors thank two anonymous reviewers and the Editor for helpful comments in earlier versions of this manuscript.

JHT and SD were supported by the Australian Research Council (FT220100629) and the Australian National Health and Medical Research Council (grant number 2017284). GB acknowledges support from the Internal Funds KU Leuven under grant agreement C14/18/094 and from the Research Foundation – Flanders ('Fonds voor Wetenschappelijk Onderzoek – Vlaanderen', G0E1420N and G098321N). The authors acknowledge efforts by originating and submitting laboratories for the sequence data in GISAID EpiCoV on which our empirical analyses are based. This research was undertaken using the LIEF HPC-GPGPU Facility hosted at the University of Melbourne. This Facility was established with the assistance of LIEF Grant LE170100200.

3 Data availability

The data underlying this article are available in GISAID at gisaid.org, and all accession numbers are provided in Supplementary Material online.

References