Title of submission to PLOS journals

John H Tay[1], Pending Pending[2,3], Sebastian Duchene[1,2,*],

1 Peter Doherty Institute for Infection and Immunity, Department of Microbiology and Immunology, University of Melbourne, Melbourne, Australia
2 Pending pending.
3 Department of Computational Biology, Institut Pasteur, Paris, France

* sduchene@pasteur.fr

Abstract

Our understanding of the evolution of many microbes has been revolutionised by the molecular clock, a statistical tool to infer evolutionary rates and timescales. The fundamental assumption of molecular clock models is that the rate at which substitutions accumulate can be described by a statistical process. In all molecular clock models evolutionary rates and times are unidentifiable, and therefore 'calibration' information is essential to obtain estimates in calendar time.

For many microbes, the sequence sampling times themselves can be often used for calibration. Phylogenetic tests of temporal signal are often used to decide whether such calibrations are reliable. Critically, in addition to the calibration information, the full Bayesian phylogenetic model also includes the molecular clock model and a branching process (tree prior). As a result, there are multiple sources of information that are difficult to untangle.

We assessed temporal signal in three microbial data sets of human and animal diseases with a range of evolutionary characteristics and with ancient DNA sequences; *Powassan virus* (POWV), the cholera bacetrium (*Vibrio cholerae*), and the syphilis bacterium (*Treponema palladium.*). We found that the tree prior can have a substantial impact on whether temporal signal is detected. To investigate this problem we conducted extensive simulations and calculated the sensitivity and specificity of these tests under several tree priors.

We find that highly informative sequence data sets are generally robust to the tree prior. In contrast, in data sets with low information content, choosing a prior that is highly informative and inconsistent with the data can result in the false rejection of temporal signal.

We propose prior sensitivity analyses and prior predictive simulations to determine whether the prior is reasonable and to improve the detection of temporal signal and maximise the information that can be drawn from molecular sequence data sets.

Author summary

Pending

Introduction

Molecular sequence data have been essential to unravel the evolutionary history of many organisms. The molecular clock is a statistical tool that posits that molecular evolution,

in the form of substitutions, follows a statistical process. For example, under the simplest molecular clock model, known as the strict clock, substitutions accumulate linearly over time along a lineage, such that the evolutionary rate is constant over time [1]. At the other end of the spectrum, relaxed molecular clocks allow every lineage in a phylogenetic tree to display a different evolutionary rate ([2] and reviewed in [3]).

All molecular clock models have a fundamental limitation, where evolutionary rates and times are unidentifiable. That is, there exist an infinite number of combinations of evolutionary rates and times that are compatible with an amount of evolutionary divergence [4, 5]. For this reason, external information, known as a molecular clock calibration is necessary to produce estimates in calendar units. As a case in point, consider two sequences whose genetic divergence from their most recent common ancestor is 10 subs/site. In the absence of calibrating information it is impossible to know *how rapidly* they evolve and *when* they diverged. The calibration can be a known evolutionary rate, such as 1 subs/site/year, or a divergence date, such as 1 year before present. The genetic distance can be divided by the evolutionary rate to infer the divergence time to infer a time to the most recent common ancestor of 10 years, or the genetic distance can be divided by the divergence date to infer the evolutionary rate, 10/subs/site/year in this case.

The finding that some organisms accumulate substitutions in a measurable timescale prompted the use of sequence sampling times for calibration [6]. The motivation behind this practice is that sequence data collected at different points in time would have accumulated a corresponding number of substitutions. With the example above, a sequence collected six months of the common ancestor would have accumulated 5 subs/site (10 subs/site/year $\times$ 0.5 years = 5 subs/site), whereas one collected after 1 year would have accrued 10 subs/site (10 subs/site/year$\times$ 1 year = 10 subs/site). As a result, sequence sampling times act as a time-calibration that is intuitively informative about the evolutionary rate.

A fundamental question to warrant the use of sampling times for molecular calibration is the extent to which a data set behaves as a measurably evolving population.

Introduction up to here. We need more information on measurably evolving populations, tests of temporal signal (including regression), and their relevance in light of genome data

$$\mathrm{P_Y} = \underbrace{H(Y_n) - H(Y_n|\mathbf{V}_n^Y)}_{S_Y} + \underbrace{H(Y_n|\mathbf{V}_n^Y) - H(Y_n|\mathbf{V}_n^{X,Y})}_{T_{X\to Y}}, \tag{1}$$

Results

Nulla mi mi, venenatis sed ipsum varius, Table 1 volutpat euismod diam. Proin rutrum vel massa non gravida. Quisque tempor sem et dignissim rutrum. Lorem ipsum dolor sit amet, consectetur adipiscing elit. Morbi at justo vitae nulla elementum commodo eu id massa. In vitae diam ac augue semper tincidunt eu ut eros. Fusce fringilla erat porttitor lectus cursus, vel sagittis arcu lobortis. Aliquam in enim semper, aliquam massa id, cursus neque. Praesent faucibus semper libero.

Pending.

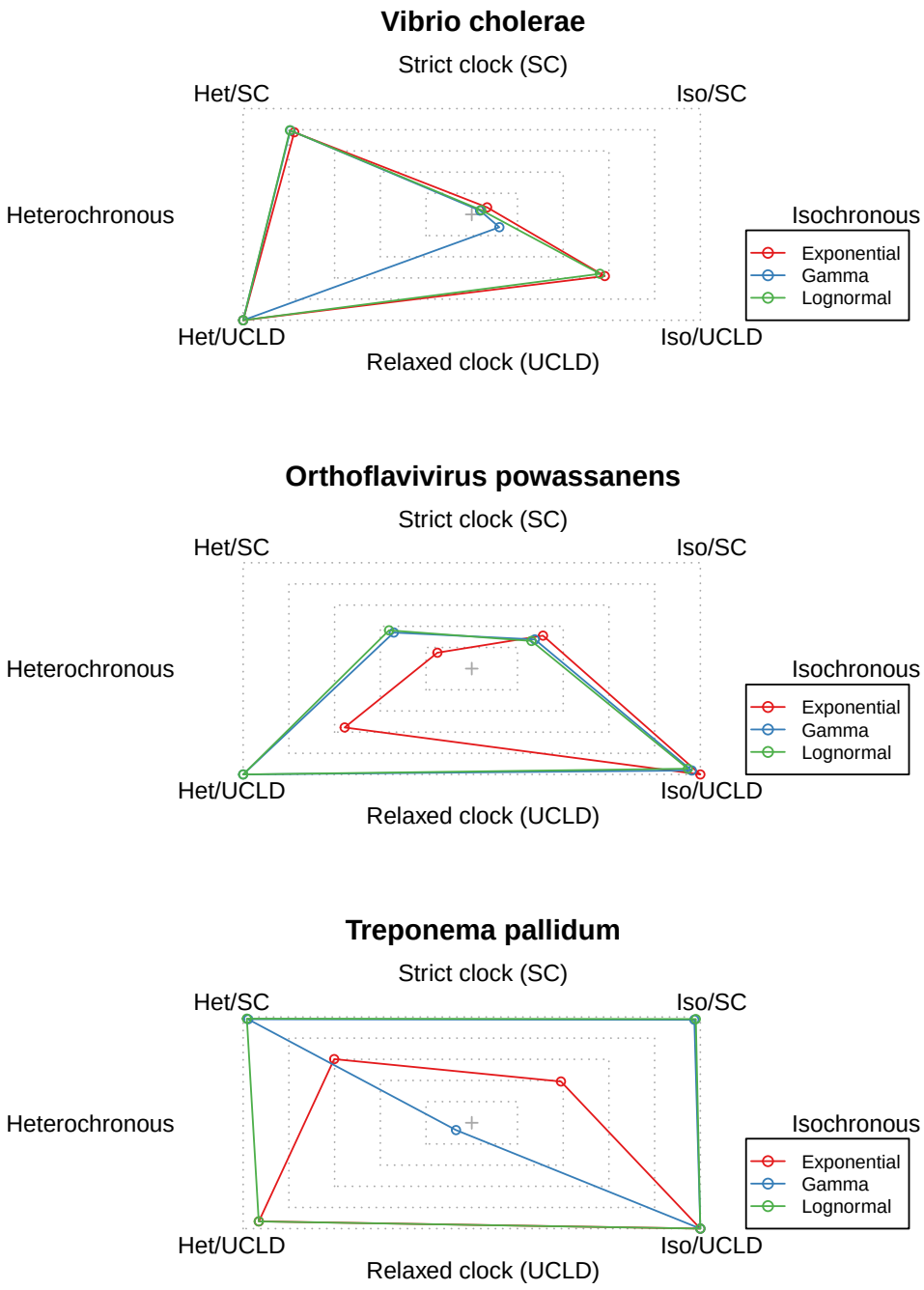

Fig 1. A polygon plot of the relative log marginal likelihoods of each microbe dataset under a different population size prior, analysed with four different configurations. Het (heterochronous) includes sampling, while iso (isochronous) does not include any sampling times. SC is strict clock and UCLD is the uncorrelated lognormal relaxed clock. Red represents an exponential hyperprior on the population size, blue is a gamma hyperprior, and green is a lognormal hyperprior.

Table 1. Table caption Nulla mi mi, venenatis sed ipsum varius, volutpat euismod diam.

Heading1				Heading2			
*cell*1*row*1	cell2 row 1	cell3 row 1	cell4 row 1	cell5 row 1	cell6 row 1	cell7 row 1	cell8 row 1
*cell*1*row*2	cell2 row 2	cell3 row 2	cell4 row 2	cell5 row 2	cell6 row 2	cell7 row 2	cell8 row 2
*cell*1*row*3	cell2 row 3	cell3 row 3	cell4 row 3	cell5 row 3	cell6 row 3	cell7 row 3	cell8 row 3

Table notes Phasellus venenatis, tortor nec vestibulum mattis, massa tortor interdum felis, nec pellentesque metus tortor nec nisl. Ut ornare mauris tellus, vel dapibus arcu suscipit sed.

LOREM and IPSUM nunc blandit a tortor

3rd level heading

Maecenas convallis mauris sit amet sem ultrices gravida. Etiam eget sapien nibh. Sed ac ipsum eget enim egestas ullamcorper nec euismod ligula. Curabitur fringilla pulvinar lectus consectetur pellentesque. Quisque augue sem, tincidunt sit amet feugiat eget, ullamcorper sed velit. Sed non aliquet felis. Lorem ipsum dolor sit amet, consectetur adipiscing elit. Mauris commodo justo ac dui pretium imperdiet. Sed suscipit iaculis mi at feugiat.

1. react
2. diffuse free particles
3. increment time by dt and go to 1

Sed ac quam id nisi malesuada congue

Nulla mi mi, venenatis sed ipsum varius, volutpat euismod diam. Proin rutrum vel massa non gravida. Quisque tempor sem et dignissim rutrum. Lorem ipsum dolor sit amet, consectetur adipiscing elit. Morbi at justo vitae nulla elementum commodo eu id massa. In vitae diam ac augue semper tincidunt eu ut eros. Fusce fringilla erat porttitor lectus cursus, vel sagittis arcu lobortis. Aliquam in enim semper, aliquam massa id, cursus neque. Praesent faucibus semper libero.

- First bulleted item.
- Second bulleted item.
- Third bulleted item.

Discussion

Nulla mi mi, venenatis sed ipsum varius, Table 1 volutpat euismod diam. Proin rutrum vel massa non gravida. Quisque tempor sem et dignissim rutrum. Lorem ipsum dolor sit amet, consectetur adipiscing elit. Morbi at justo vitae nulla elementum commodo eu id massa. In vitae diam ac augue semper tincidunt eu ut eros. Fusce fringilla erat porttitor lectus cursus, vel sagittis arcu lobortis. Aliquam in enim semper, aliquam massa id, cursus neque. Praesent faucibus semper [6] libero [**?**].

Materials and methods

Etiam eget sapien nibh

Nulla mi mi, Fig 2 venenatis sed ipsum varius, volutpat euismod diam. Proin rutrum vel massa non gravida. Quisque tempor sem et dignissim rutrum. Lorem ipsum dolor sit amet, consectetur adipiscing elit. Morbi at justo vitae nulla elementum commodo eu id massa. In vitae diam ac augue semper tincidunt eu ut eros. Fusce fringilla erat porttitor lectus cursus, S1 Video vel sagittis arcu lobortis. Aliquam in enim semper, aliquam massa id, cursus neque. Praesent faucibus semper libero.

Fig 2. Bold the figure title. Figure caption text here, please use this space for the figure panel descriptions instead of using subfigure commands. A: Lorem ipsum dolor sit amet. B: Consectetur adipiscing elit.

Supporting information

S1 Fig. Bold the title sentence. Add descriptive text after the title of the item (optional).

S2 Fig. Lorem ipsum. Maecenas convallis mauris sit amet sem ultrices gravida. Etiam eget sapien nibh. Sed ac ipsum eget enim egestas ullamcorper nec euismod ligula. Curabitur fringilla pulvinar lectus consectetur pellentesque.

S1 File. Lorem ipsum. Maecenas convallis mauris sit amet sem ultrices gravida. Etiam eget sapien nibh. Sed ac ipsum eget enim egestas ullamcorper nec euismod ligula. Curabitur fringilla pulvinar lectus consectetur pellentesque.

S1 Video. Lorem ipsum. Maecenas convallis mauris sit amet sem ultrices gravida. Etiam eget sapien nibh. Sed ac ipsum eget enim egestas ullamcorper nec euismod ligula. Curabitur fringilla pulvinar lectus consectetur pellentesque.

S1 Appendix. Lorem ipsum. Maecenas convallis mauris sit amet sem ultrices gravida. Etiam eget sapien nibh. Sed ac ipsum eget enim egestas ullamcorper nec euismod ligula. Curabitur fringilla pulvinar lectus consectetur pellentesque.

S1 Table. Lorem ipsum. Maecenas convallis mauris sit amet sem ultrices gravida. Etiam eget sapien nibh. Sed ac ipsum eget enim egestas ullamcorper nec euismod ligula. Curabitur fringilla pulvinar lectus consectetur pellentesque.

Acknowledgments

Cras egestas velit mauris, eu mollis turpis pellentesque sit amet. Interdum et malesuada fames ac ante ipsum primis in faucibus. Nam id pretium nisi. Sed ac quam id nisi malesuada congue. Sed interdum aliquet augue, at pellentesque quam rhoncus vitae.

References

1. Zuckerkandl E, Pauling L. Evolutionary divergence and convergence in proteins. In: Evolving genes and proteins. Elsevier; 1965. p. 97–166.

2. Drummond AJ, Ho SYW, Phillips MJ, Rambaut A. Relaxed phylogenetics and dating with confidence. PLoS Biology. 2006;4(5):e88.

3. Ho SY, Duchêne S. Molecular-clock methods for estimating evolutionary rates and timescales. Molecular Ecology. 2014;23(24):5947–5965.

4. Yang Z, Rannala B. Bayesian estimation of species divergence times under a molecular clock using multiple fossil calibrations with soft bounds. Molecular biology and evolution. 2006;23(1):212–226.

5. Dos Reis M, Yang Z. The unbearable uncertainty of Bayesian divergence time estimation. Journal of Systematics and Evolution. 2013;51(1):30–43.

6. Drummond AJ, Pybus OG, Rambaut A, Forsberg R, Rodrigo AG. Measurably evolving populations. Trends in ecology & evolution. 2003;18(9):481–488.