Bioinformatics, 2016, 1–5
doi: 10.1093/bioinformatics/btw421
Advance Access Publication Date: 13 July 2016
Discovery note

OXFORD

Phylogenetics

Estimating evolutionary rates using time-structured data: a general comparison of phylogenetic methods

Sebastián Duchêne,[1,2,3,*,†] Jemma L. Geoghegan,[1,2,†] Edward C. Holmes[1,2] and Simon Y.W. Ho[2]

[1]Marie Bashir Institute of Infectious Diseases and Biosecurity, Charles Perkins Centre, Sydney Medical School, University of Sydney, Sydney, NSW 2006, Australia, [2]School of Life and Environmental Sciences, University of Sydney, Sydney, NSW 2006, Australia and [3]Centre for Systems Genomics, University of Melbourne, Parkville, Victoria 3010, Australia

*To whom correspondence should be addressed.
†The authors wish it to be known that, in their opinion, the first two authors should be regarded as Joint First Authors.
Associate Editor: Janet Kelso

Received on March 21, 2016; revised on June 2, 2016; accepted on June 27, 2016

Abstract

Motivation: In rapidly evolving pathogens, including viruses and some bacteria, genetic change can accumulate over short time-frames. Accordingly, their sampling times can be used to calibrate molecular clocks, allowing estimation of evolutionary rates. Methods for estimating rates from time-structured data vary in how they treat phylogenetic uncertainty and rate variation among lineages. We compiled 81 virus data sets and estimated nucleotide substitution rates using root-to-tip regression, least-squares dating and Bayesian inference.

Results: Although estimates from these three methods were often congruent, this largely relied on the choice of clock model. In particular, relaxed-clock models tended to produce higher rate estimates than methods that assume constant rates. Discrepancies in rate estimates were also associated with high among-lineage rate variation, and phylogenetic and temporal clustering. These results provide insights into the factors that affect the reliability of rate estimates from time-structured sequence data, emphasizing the importance of clock-model testing.

Contact: sduchene@unimelb.edu.au or garzonsebastian@hotmail.com

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

Reliable inference of molecular evolutionary time-scales depends on accurate estimates of substitution rates. The simplest way to estimate these rates is to assume that the divergence of nucleotide or amino acid sequences occurs constantly over time (Zuckerkandl and Pauling, 1962), a model known as the strict molecular clock. Under this model, the rate of evolution corresponds to the amount of sequence divergence between a pair of lineages, divided by the time to their most recent common ancestor. For data sets in which appreciable genetic change has accumulated during the window of sampling, such that the sequence data are time structured, the sequence ages can be used as calibrations to estimate the rate (reviewed by Rieux and Balloux, 2016). Sequence data from rapidly evolving pathogens and ancient specimens often meet these conditions (Biek *et al.*, 2015; Drummond *et al.*, 2003).

There has been a surge in phylogenetic methods that can explicitly handle time-structured sequence data. These methods include root-to-tip regression (implemented in TempEst; Rambaut *et al.*, 2016), least-squares dating (implemented in LSD; To *et al.*, 2016), and Bayesian phylogenetic inference (e.g. BEAST; Drummond *et al.*,

2012). These methods differ in how they treat phylogenetic uncertainty and in their assumptions about rate variation among lineages.

Root-to-tip regression uses a phylogenetic tree with branch lengths in units of genetic divergence (e.g. substitutions/site). A regression of the root-to-tip number of expected substitutions as a function of the sampling times is fitted (Korber *et al.*, 2000). The slope of the regression corresponds to the rate, the x-intercept is the time to the most recent common ancestor, and the R^2 value can be used as a measure of clocklike behavior. Crucially, however, the data points in the regression are not phylogenetically independent (Rambaut *et al.*, 2016). In particular, deep branches in the tree contribute to multiple root-to-tip distances, leading to pseudoreplication. Least-squares dating also requires a phylogenetic tree. Instead of fitting a regression, it estimates the rate using a normal approximation of the Langley-Fitch algorithm (Langley and Fitch, 1974). Both root-to-tip regression and least-squares dating assume clock-like evolution and use rooted trees, but they are able to estimate the position of the root as part of the analysis. However, the two methods do require a fixed tree, such that phylogenetic uncertainty is not explicitly taken into account, although this can be done indirectly by conducting the analysis over a set of bootstrap trees.

Bayesian approaches to estimating evolutionary rates require a model of rate variation to be specified. In most Bayesian dating programs, such as BEAST, several clock models are available. The strict clock (SC) model assumes rate homogeneity throughout the tree. In contrast, relaxed-clock models allow a different rate along each branch in the tree, although these are assumed to follow a chosen statistical distribution. For example, branch rates can be treated as independent and identically distributed samples from a lognormal distribution or an exponential distribution in the uncorrelated lognormal (UCLD) and uncorrelated exponential (UCED) relaxed-clock models, respectively (Drummond *et al.*, 2006). An advantage of Bayesian methods is that the rate estimate is integrated over the uncertainty of all of the other parameters, including the phylogenetic tree.

The fundamental differences among the three methods might be expected to result in different estimates of evolutionary rates (Fourment and Holmes, 2014). To test for congruence in rate estimates we collected sequence data sets for 81 different RNA and DNA viruses. For each data set we estimated the evolutionary rate using root-to-tip regression in TempEst v1.5 (Rambaut *et al.*, 2016), least-squares dating in LSD v0.2 (To *et al.*, 2016), and Bayesian phylogenetic inference in BEAST v1.8 (Drummond *et al.*, 2012). We found some discrepancies in the estimates, most of which were explained by high-among lineage rate variation, and phylogenetic and temporal clustering. As such, these factors should be routinely assessed in empirical data to improve estimates of evolutionary rates and time scales.

2 Methods

We obtained 81 nucleotide sequence virus data sets from GenBank (available at https://github.com/sebastianduchene/virus_rate_variation). Each data set contained between 9 and 120 sequences, with alignments ranging from 305 to 10,066 nucleotides. All data sets included the sampling times (calendar dates) and were sampled over time-frames ranging from ~0.5 to 86 years.

We estimated the substitution rate for each data set using three different methods (TempEst, LSD, and BEAST). For all methods we used the sampling times for calibration. TempEst and LSD both require a phylogenetic tree. We used PhyML v3.1 (Guindon *et al.*, 2010) to estimate the tree topology and branch lengths using maximum likelihood, under the GTR + Γ substitution model. We generated a root-to-tip regression for each data set in TempEst, where the root of the tree was selected to maximize the R^2. We also used LSD with the same maximum likelihood trees as above. For this method we set the minimum threshold for the rate at 10^{-10} substitutions/site/year, and the position of the root was inferred in the analysis.

We analyzed the data in BEAST using the same substitution model as in our maximum likelihood analyses. For the substitution rate we used a uniform prior with 0 and 1 as the minimum and maximum bounds, respectively, and the constant-size coalescent tree prior. We set the chain length to 10^7 steps, sampling every 5000, and verifying that the effective sample size of each parameter was at least 200. TempEst and LSD always assume clock-like evolution, although the former is able to tolerate some stochastic rate heterogeneity (To *et al.*, 2016). In BEAST, however, we analyzed the data using both the SC and relaxed (UCLD and UCED) clock models. To compare the estimates from BEAST with those from TempEst and LSD, we chose between the SC and UCLD models by inspecting the coefficient of rate variation in the UCLD model. This parameter is the standard deviation of the branch rates divided by the mean rate (Drummond *et al.*, 2006). If the lower 95% limit of the posterior density is close to zero, the data can be considered to have evolved in a clock-like manner. In such cases we report the rate estimates from the SC model, otherwise we report those from the UCLD model. Importantly, this method for clock model selection has similar performance to more computationally intensive methods (Ho *et al.*, 2015). To determine the impact of the choice of clock model in the estimates from BEAST we also compared the SC and UCLD estimates with those obtained from the UCED model.

One important requirement of Bayesian estimates of evolutionary rates and time scales is that the data should have sufficient temporal structure. We assessed whether this was the case for our data sets by conducting a date-randomization test (Ramsden *et al.*, 2009). This test consists of randomizing the sampling times of the sequences and repeating the analysis. If the data have sufficient temporal structure the rate estimate obtained using the correct sampling times should be different from those based on the date-randomized replicates. We used two criteria for this test: CR1 and the more stringent CR2 (Duchêne *et al.*, 2015). Under CR1, the data have sufficient temporal structure if the mean estimate based on the correct sampling times is not contained within the 95% credible interval of the rate estimate from any of the date-randomized replicates. For CR2, the 95% credible interval of the estimate based on the correct sampling times should not overlap with those from the date-randomized replicates. Importantly, we used this test for the rate estimates from BEAST, where it is typically performed, and not for TempEst or LSD which produce point estimates.

We considered three factors that might lead to discrepancies in the estimates between methods: topological incongruence, phylogenetic and temporal clustering, and among-lineage rate variation. To measure topological incongruence we computed the topological distance between the maximum likelihood and the Bayesian highest-clade-credibility trees using the PH85 metric (Penny and Hendy, 1985). To allow comparison between data sets, we standardized the PH85 distance by the number of taxa in each data set.

Phylogenetic and temporal clustering occurs when sequences with similar sampling times are closely related in the phylogenetic tree, a pattern that leads to unreliable rate estimates and poor performance of the date-randomization test (Duchêne *et al.*, 2015; Murray *et al.*, 2015). To quantify phylogenetic and temporal clustering, we assessed the correlation between phylogenetic distance

and sampling times in the maximum likelihood trees. Our measure of phylogenetic distance is based on the number of nodes separating two tips. For every pair of tips, we calculated the difference in sampling times and their phylogenetic distance, and calculated a Pearson's correlation coefficient (ρ). To obtain a *P*-value we generated a null distribution of ρ by randomizing the sampling times in the trees 100 times. This procedure is appropriate for our data because the samples are not phylogenetically independent and it does not make parametric assumptions. A *P*-value of <0.05 means that there is a significant association between sampling times and phylogenetic distance.

3 Results

For 28 of the 81 data sets it was not possible to estimate the rate using TempEst or LSD. This occurs when the rate estimate from TempEst is negative, or when that from LSD corresponds to the lower threshold value. We found that 21 of these data sets displayed considerable rate variation across branches, with coefficients of rate variation that were distinctly non-zero (Supplementary Fig. S1).

We therefore focus on the 53 data sets for which we could estimate the rate using all three methods (Supplementary Table S1). We also calculated the regression *P*-value and found that 16 data sets yielded positive slopes that were non-significant. However, since the *P*-value is difficult to interpret in this context due to pseudoreplication, we retained these data sets in our analysis.

We plotted a pairwise comparison of the rates from the different methods. We measured their median proportional difference between the estimates with sufficient temporal structure according to CR2 in the date-randomization test. For example, for TempEst vs. LSD the proportional difference is the estimate from TempEst minus that from LSD, divided by that from TempEst. For the estimates obtained with BEAST we report the mean, while for those from TempEst and LSD we use the point estimates. To measure bias we noted the percentage of data sets for which the estimates along the *y*-axis are larger than those on the *x*-axis. Although TempEst tended to produce higher estimates than LSD, with a bias of more than 80%, the estimates from TempEst were only about 0.14 times higher those for LSD measured using the median proportional difference (Table 1 and Fig. 1a). In comparison, the estimates from BEAST were less congruent with those from both LSD and TempEst, with median proportional differences of −0.40 and −0.25, respectively.

Although the estimates from BEAST were not systematically higher or lower than those from LSD and TempEst, the level of bias was greater for data sets with considerable among-lineage rate variation. For example, in our comparison of BEAST versus LSD (Fig. 1c), 83.3% of data sets analyzed using the UCLD model had higher rates than those from LSD, but only 27.3% of those analyzed with the SC model displayed this pattern (Table 1). This is likely because LSD and TempEst assume a strict molecular clock, whereas BEAST can use relaxed-clock models. This result is also evident in our comparison of the different clock models in BEAST. The 95% credible intervals of estimates from the three clock models largely overlap and there is broad agreement between the estimates using the UCLD and UCED models. However, the mean values from the two relaxed-clock models are nearly always higher than those from the SC model (Table 2; Fig. 2). A probable reason for this pattern is that estimates from relaxed-clock models have higher uncertainty than those from the SC model, which might be upwardly biased

Table 1. Median proportional differences and bias for comparisons of three different methods for estimating substitution rates from time-structured sequence data

	TempEst versus LSD		BEAST versus TempEst		BEAST versus LSD	
	Proportional difference	Bias	Proportional difference	Bias	Proportional difference	Bias
Complete data set	0.14	81.5	−0.25	34.5	−0.40	64.0
UCLD	0.28	77.8	0.15	48.0	0.48	83.3
SC	−0.04	83.4	−0.84	10.0	−1.17	27.3

The proportional difference for two methods, such as TempEst versus LSD, is the difference between their estimates divided by the estimate from TempEst. The measure of bias is the percentage of estimates on the *y*-axis that are larger than those on the *x*-axis (Fig. 1).

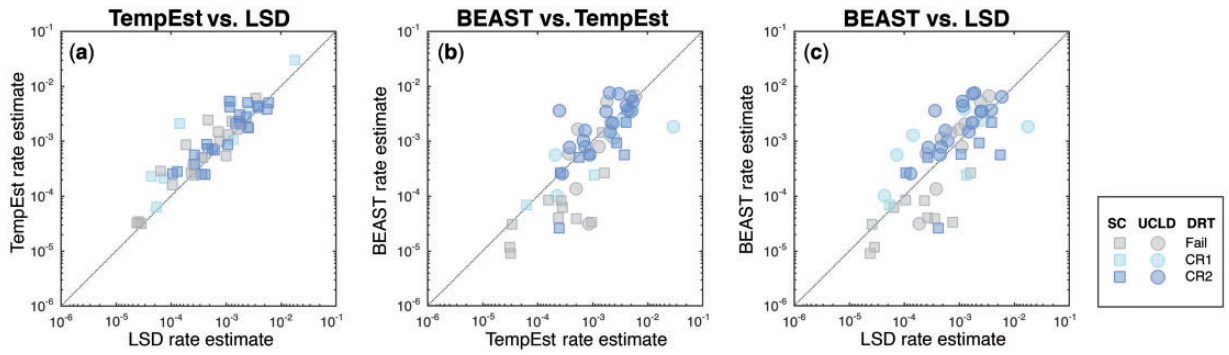

Fig. 1. Estimates of nucleotide substitution rate (substitutions/site/year) for all data sets using root-to-tip regression in TempEst, least-squares dating in LSD, and Bayesian phylogenetic inference in BEAST, and plotted against each other. **(a)** TempEst versus LSD; **(b)** BEAST versus TempEst; and **(c)** BEAST versus LSD. For the results from BEAST, the data points correspond to the mean rate estimates. Data points are colored according to the results of the date-randomization test conducted for the Bayesian estimates only: passing CR2 (dark blue); passing CR1 (light blue); and failing both criteria (grey). Circular markers indicate that the UCLD clock was implemented, whereas square markers denote estimates made using a SC. The dotted line represents $y = x$, denoting equality between the rate estimates from the two methods

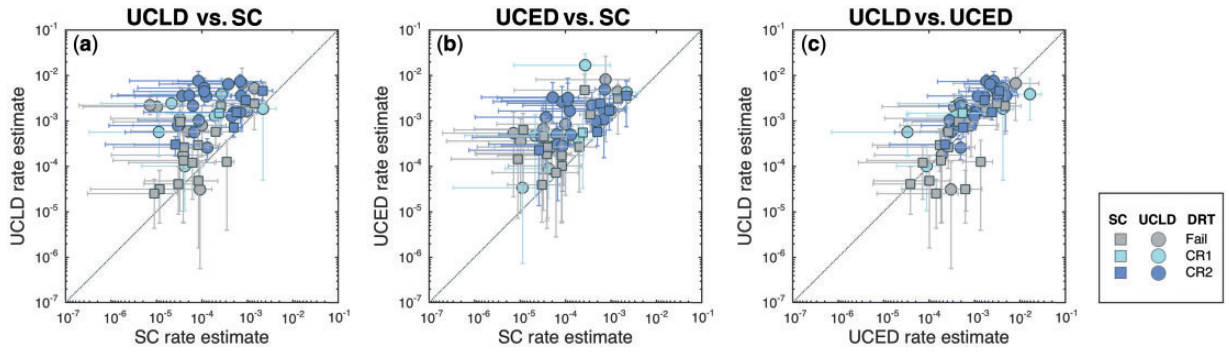

Fig. 2. Pairwise comparisons of mean estimates and 95% credible intervals of nucleotide substitution rate (substitutions/site/year) using Bayesian phylogenetic inference in BEAST. Estimates were made using a SC, UCLD clock, or UCED clock. **(a)** UCLD versus SC; **(b)** UCED versus SC; and **(c)** UCLD versus UCED. The colors and shapes of the data points correspond to those in Figure 1

Table 2. Median proportional distance, bias, and percentage of overlapping 95% credible intervals for rate estimates using the SC, UCLD and UCED clock models in BEAST

	UCLD versus SC	UCED versus SC	UCLD versus UCED
Distance	0.12	0.70	−0.04
Bias	76.2	100.0	71.0
Overlapping	68.3	89.2	86.5

Measures of distance and bias are the same as those in Table 1.

because the rate naturally has a lower bound at zero. These factors can lead to a higher mean value (Firth *et al.*, 2010; Ho *et al.*, 2007). Consequently, the use of relaxed molecular clocks can lead to overestimates of the mean substitution rate when the data in fact fit a SC, indicating that clock-model testing should be routinely performed.

We looked for common features among those data sets for which the methods analyzed here produced very different rate estimates. Notably, these data sets typically displayed high among-lineage rate variation, incongruence between the maximum-likelihood and Bayesian estimates of tree topology, and phylogenetic and temporal clustering. However, high among-lineage rate variation was the only factor to display a significant (i.e. $P < 0.05$) and positive correlation with the degree of incongruence across all data sets (Supplementary Table S1; Supplementary Fig. S2). One example of a data set with incongruent rate estimates among methods and which displayed all these factors is the Hantaan virus data set, with an especially high coefficient of rate variation (mean 1.29, 95% credible interval 0.71–2.00), very strong phylogenetic and temporal clustering ($P < 0.05$), and high relative topological distance between the maximum-likelihood and Bayesian trees (1.08). Strikingly, the absolute sampling time-frame does not appear to predict whether there would be conflict between the rate estimates from different methods.

Overall, our study demonstrates that methods for analyzing time-structured sequence data largely produce congruent estimates of substitution rates, provided that clock-model testing is performed and that the data meet certain criteria. In practice, therefore, it is important to verify that the data have sufficient temporal structure and have no phylogenetic or temporal clustering. If the data display phylogenetic and temporal clustering, a modification of the date randomization test can be conducted (Duchêne *et al.*, 2015; Murray *et al.*, 2015; Rieux and Balloux, 2016), but this is not trivial for data sets without discrete monophyletic groups with the same sampling times, as in our data. The performance of these methods should also be investigated using analyses of other time-structured data sets, such as those from bacteria and ancient DNA. These data sets form the basis of a wide range of important evolutionary studies, but pose substantially different analytical challenges from the virus data sets studied here.

Funding

This research was funded by a McKenzie Fellowship from the University of Melbourne awarded to S.D., a Judith and David Coffey fellowship from the Charles Perkins Centre, University of Sydney, awarded to J.L.G., an NHMRC Australia Fellowship (AF30) awarded to E.C.H., and an Australian Research Council fellowship awarded to S.Y.W.H.

Conflict of Interest: none declared.

References

Biek,R. *et al.* (2015) Measurably evolving pathogens in the genomic era. *Trends Ecol. Evol.*, **30**, 306–313.

Drummond,A.J. *et al.* (2006) Relaxed phylogenetics and dating with confidence. *PLOS Biol.*, **4**, e88.

Drummond,A.J. *et al.* (2003) Measurably evolving populations. *Trends Ecol. Evol*, **18**, 481–488.

Drummond,A.J. *et al.* (2006) Relaxed phylogenetics and dating with confidence. *PLOS Biol*, **4**, 699–710.

Duchêne,S. *et al.* (2015) The performance of the date-randomization test in phylogenetic analyses of time-structured virus data. *Mol. Biol. Evol.*, **32**, 1895–1906.

Firth,C. *et al.* (2010) Using time-structured data to estimate evolutionary rates of double-stranded DNA viruses. *Mol. Biol. Evol.*, **27**, 2038–2051.

Fourment,M. and Holmes,E.C. (2014) Novel non-parametric models to estimate evolutionary rates and divergence times from heterochronous sequence data. *BMC Evol. Biol.*, **14**, 163.

Guindon,S. *et al.* (2010) New algorithms and methods to estimate maximum likelihood phylogenies: assessing the performance of PhyML 3.0. *Syst. Biol.*, **59**, 307–321.

Ho,S.Y.W. *et al.* (2007) Evidence for time dependency of molecular rate estimates. *Syst. Biol.*, **56**, 515–522.

Ho,S.Y.W. *et al.* (2015) Simulating and detecting autocorrelation of molecular evolutionary rates among lineages. *Mol. Ecol. Resour.*, **15**, 688–696.

Korber,B. *et al.* (2000) Timing the ancestor of the HIV-1 pandemic strains. *Science*, **288**, 1789–1796.

Langley,C.H. and Fitch,W.M. (1974) An examination of the constancy of the rate of molecular evolution. *J. Mol. Evol.*, **3**, 161–177.

Murray,G.G.R. *et al.* (2015) The effect of genetic structure on molecular dating and tests for temporal signal. *Methods Ecol. Evol.*, **7**, 80–89.

Penny,D. and Hendy,M. (1985) The use of tree comparison metrics. *Syst. Zool.*, **34**, 75–82.

Rambaut,A. *et al.* (2016) Exploring the temporal structure of heterochronous sequences using TempEst (formerly Path-O-Gen). *Virus Evol.*, **2**, vew007.

Ramsden,C. *et al*. (2009) Hantavirus evolution in relation to its rodent and insectivore hosts: no evidence for codivergence. *Mol. Biol. Evol.*, **26**, 143–153.

Rieux,A. and Balloux,F. (2016) Inferences from tip-calibrated phylogenies: a review and a practical guide. *Mol. Ecol.*, **25**, 1911–1924.

To,T.H. *et al*. (2016) Fast dating using least-squares criteria and algorithms. *Syst. Biol.*, **65**, 82–97.

Zuckerkandl,E. and Pauling,L. (1962) Molecular disease, evolution and genic heterogeneity. In, Kasha,M. and Pullman,B. (eds) *Horizons in Biochemistry*. Academic press, New York, pp. 189–225.