Bioinformatics
doi.10.1093/bioinformatics/xxxxxx
Advance Access Publication Date: Day Month Year
Application Note

OXFORD

node.dating: Timing ancestral dates in phylogenetic trees in R

Bradley R. Jones [1, 2,*] and Art F. Y. Poon [2,3]

[1]Faculty of Health Sciences, Simon Fraser University, Burnaby, V5A 1S6, Canada,
[2]BC Centre for Excellence in HIV/AIDS, Vancouver, V6Z 1Y6, Canada and
[2]Department of Medicine, University of British Columbia, V5Z 1M9, Canada

*To whom correspondence should be addressed.

Associate Editor: XXXXXXX

Received on XXXXX; revised on XXXXX; accepted on XXXXX

Abstract

Summary: Phylogenetic trees encode the evolutionary distances between species or populations. With sufficient information, these evolutionary distances can be rescaled over time to provide estimates of the dates of the most recent ancestors of the species. Here we present the R software `node.dating`, which uses a maximum likelihood method to estimate the dates of the internal nodes of a phylogenetic tree.
Availability and Implementation: `node.dating` is written in R and requires the R package, *APE*. `node.dating` is available ()
Contact: brj1@sfu.ca
Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

Phylogenetic trees represent the evolutionary relationships among populations or species through their common ancestors. The length of a branch in the phylogeny corresponds to the expected amount of evolution between the ancestor and its descendant, where the passage of time and the rate of evolution are confounded. However, when there is external information available on the location of nodes in the tree in time, then the branch lengths can be rescaled with respect to time given sufficient variation in node timings for measurable evolution to occur. Thus, the internal nodes of a time-scaled tree estimate the dates that the respective lineages diverged from their common ancestor (Kumar and Hedges, 2016). These date estimates are an important resource for reconstructing the evolutionary history of species (Shapiro et al., 2004). In molecular epidemiology, these dates can also provide a rough approximation of transmission times during an outbreak of infectious disease (Ypma et al., 2013).

A multitude of software has been developed to date MRCA's and create time trees using various techniques such as: linear regression (Rambaut et al., 2016), maximum likelihood (Rambaut, 2000; Sanderson, 2003; Yang, 2007), Bayesian analysis (Drummond and Rambaut, 2007), heuristics (Drummond and Rodrigo, 2000; Yang et al., 2007), and least squares (To et al., 2016). However, most of this software is difficult to access or use. Our software, `node.dating`, uses a maximum likelihood approach to date the internal nodes of a phylogenetic tree. `node.dating` is written in *R* and is compatible with the *R* package *APE* (Paradis et al., 2004); it is available (). *APE* has the capability to estimate the dates of internal nodes via the `chronos` function; however `chronos` does not incorporate the dates of the tips into its solution and thus does not offer the same precision as `node.dating`.

2 Algorithm

We start with a rooted phylogenetic tree with edge length equal to genetic distance and the dates when each tip of the tree was sampled. A simple linear regression is used to estimate the mutation rate assuming a strict molecular clock. To estimate the dates of the internal nodes, we follow an approach given by Felsenstein (1981) and inspired by *TipDates* (Rambaut, 2000). A maximum likelihood method is applied locally to date each internal node. Then using these estimates, the algorithm iterates until a sufficiently approximate maximum likelihood solution is obtained.

3 Runtime and accuracy

In order to test the viability of `node.dating`, we ran `node.dating` on 50 simulated trees. We evaluated the results of `node.dating` using a weighted root mean squared error (RMSE). The simulation method and weighted RMSE are detailed in the Supplementary Information.

We compared the results of `node.dating` of the 50 simulated trees against a modified version of *TempEst*'s (formerly *Path-O-Gen*) internal

Table 1. The runtime and accuracy of `node.dating` and comparable software on 50 simulated trees

Software	Runtime (s)	RMSE (days)
`node.dating` (initial)	1.64	29.2
`node.dating` (1 step)	2.26	26.7
`node.dating` (10 steps)	7.62	23.7
`node.dating` (100 steps)	61.8	22.8
`node.dating` (1000 steps)	596	22.6
`node.dating` (10^4 steps)	5940	22.1
TempEst	11.2	132
LSD	0.707	24.9
BEAST (10^4 steps)	181	419
BEAST (10^6 steps)	6840	20.1

node dating, *LSD* using a weighted root mean squared error (RMSE). We also compared our results using *BEAST*, but on the simulated sequences instead of their reconstructed phylogenies. *TempEst* uses the prediction given by a linear regression to estimate the dates of the internal nodes and *LSD* uses a least squares method.

The average RMSE's of *TempEst* and *LSD* are higher than the average weighted RMSE of the MRCA using `node.dating` with 1000 steps, though the weighted RMSE of *LSD* is comparable to the RMSE of `node.dating`. However, the main purpose of *TempEst* is not to date internal nodes, but to detect/verify the presence of a strict molecular clock. We did not use *TempEst* or *LSD*'s root-to-tip regression in our experiments because we merely wanted to compare the internal node dating.

BEAST performed the best of all methods when running 10^6 steps, but this also took the most time. With 10^4 steps *BEAST* performed terribly and in around the same amount of time as `node.dating` at 100 steps. However, the results using *BEAST* are not completely comparable since we used the sequence data for *BEAST* instead of the reconstructed phylogenies.

4 Visualization

In this section, we exhibit a visualization of sequence data using the estimated dates of internal nodes. We retrieved intra-host patient-derived sequences from Patient 16617 on the LANL HIV database (LANL, accessed June 24, 2015). This patient's sequence data was collected in Llewellyn et al. (2006) as Patient 1180. We reconstructed the sequences' phylogenetic tree using the method described in the Supplementary Information. We then estimated the dates of the internal nodes using `node.dating` with 1000 iterations. Finally we plotted the genetic distance from the root versus time of the internal nodes and the sampled sequences. This plot displayed shown in Figure 1.

5 Future work

One drawback of our methodology is that it assumes that the phylogeny follows a strict molecular clock. However, the local likelihood model can be extended to incorporate a variable molecular clock; future work would like to include this extension. The molecular clock assumption also implies that mutations are strictly additive over time, which is not true. It may also be possible to incorporate this "negative" evolution into the model.

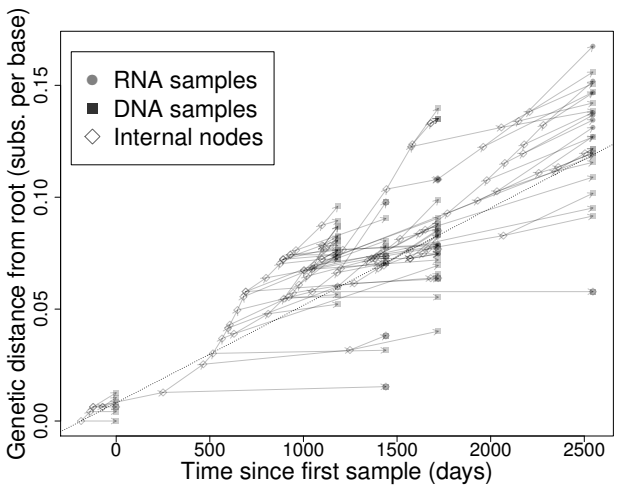

Fig. 1. Genetic distance from the root versus time of the sequences from Patient 16617 from the LANL HIV database. The dates of the internal nodes were estimated using `node.dating` and the internal nodes are included in the plot. The edges between adjacent nodes of the tree are drawn as solid lines and the molecular clock is drawn by a dashed line.

Acknowledgements

We would like to thank Richard Liang for his aid in automating *TempEst*.

Funding

This work was supported by the Canadian Institutes of Health Research (CIHR Team Grant: HIV Cure Research — The Canadian HIV Cure Research Enterprise; CanCure), by the Bill and Melinda Gates Foundation Award Number OPP1110049, and by a CIHR operating grant to Art Poon (HOP-111406).

References

Drummond, A. and Rodrigo, R., A G (2000) Reconstructing genealogies of serial samples under the assumption of a molecular clock using serial-sample UPGMA. *Molecular Biology and Evolution*, 17(12), 1807–1815.

Drummond, A. J. and Rambaut, A. (2007) BEAST: Bayesian evolutionary analysis by sampling trees. *BMC Evolutionary Biology*, 7, 214.

Felsenstein, J. (1981) Evolutionary trees from DNA sequences: a maximum likelihood approach. *Journal of Molecular Evolution*, 17, 368–376.

Kumar, S. and Hedges, S. B. (2016) Advances in time estimation methods for molecular data. *Molecular Biology and Evolution*, 33(4), 863–869.

LANL (accessed June 24, 2015) Los Alamos HIV Database. URL `http://www.hiv.lanl.gov/`.

Llewellyn, N. et al. (2006) Continued evolution of HIV-1 circulating in blood monocytes with antiretroviral therapy: genetic analysis of HIV-1 in monocytes and CD4+ T cells of patients with discontinued therapy. *J Leukoc Biol*, 80(5), 1118–1126.

Paradis, E. et al. (2004) APE: analyses of phylogenetics and evolution in R language. *Bioinformatics*, 20(2), 289–290.

Rambaut, A. (2000) Estimating the rate of molecular evolution: incorporating non-contemporaneous sequences into maximum likelihood phylogenies. *Bioinformatics*, 16(4), 395–399.

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

Sanderson, M. J. (2003) r8s: inferring absolute rates of molecular evolution and divergence times in the absence of a molecular clock. *Bioinformatics*, 19(2), 301–302.

Shapiro, B. et al. (2004) Rise and fall of the Beringian Steppe bison. *Science*, 306, 1561–1565.

To, T.-H. et al. (2016) Fast dating using least-squares criteria and algorithms. *Systematic Biology*, 65(1), 82–97.

Yang, Y. (2007) PAML 4: phylogenetic analysis by maximum likelihood. *Molecular Biology and Evolution*, 24(8), 1585–1591.

Yang, Z. et al. (2007) Tree and rate estimation by local evaluation of heterochronous nucleotide data. *Bioinformatics*, 23(2), 169–176.

Ypma, R. J. F. et al. (2013) Relating phylogenetic trees to transmission trees of infectious disease outbreaks. *Genetics*, 195(3), 1055–1062.