Should protracted speciation be incorporated in phylogenetic tree construction methods?

Richèl J.C. Bilderbeek & Rampal S. Etienne

July 5, 2016

1 Abstract

The construction of phylogenies has proven invaluable in answering evolutionary biological questions. Our current phylogenetic tools ignore the fact that speciation takes time, which has an effect unknown in phylogeny reconstruction. Here, we measure the errors that the Bayesian phylogenetic software tool BEAST2 gives when recovering simulated phylogenies, for different times-to-speciate, under a range of additional parameter settings. It has been found that branch lengths are consistently and strongly underestimated for biologically relevant parameters. This research shows that protractedness is a complexity of nature that should not be ignored and should be incorporated in our phylogenetic tools.

2 Introduction

Speciation takes time Although we know that speciation takes time, we commonly ignore this when constructing a phylogeny, by chosing a constant-rate birth-death model as a speciation model. The constant-rate birth-death model (as described in for example [7]) is among the simplest speciation models, and assumes instant speciation and extinction rates, captured by its two parameters. The constant-rate birth-death model is popular for its simplicity, yet has also served as a starting point for more elaborate specation models.

Other non-protracted speciation models Other speciation models may assume that speciation rate changes in time [11], is dependent on the amount of species present [3], or is trait dependent [6]. The original model by making speciation rate dependent on time, diversity or trait value. Also these extended models assume speciation is instantaneous.

Protracted speciation model The protracted speciation model [12] allows for speciation taking time. It adds an additional species state (see also figure 1): the 'incipient' stage, which a lineages has to complete before becoming a 'good' species. The protracted speciation model makes no assumption about the exact biolological mechanism by which reproductive isolation is acquired [5, 4, 12], yet matches best to the peripatric mode of speciation, in which species occupy a new niche, without gene flow between parental and this incipient species. The protracted speciation model makes no assumption about speciation and extinction rates (these may be contant or depend on time, diversity or trait value), but these features can easily be added to the model.

Use of a speciation model in inferring a phylogeny Speciation models are widely used to make inferences from genetic data. There are multiple computer programs to create phylogenies and/or parameter estimates from aligned DNA sequences. BEAST2 is a widely used tool that allows for a Bayesian approach to phylogenetics[1]. BEAST2 supplies the user with multiple speciation models, that all assume instantaneous speciation.

This study This simulation investigates the consequence of BEAST2 using instantaneous speciation, by creating a 'true' tree that is protracted and seeing how well BEAST2 can recover it. It is expected that for higher protractedness (thus deviation from BEAST2 its assumptions), the error will increase. It is unknown, however, if and when this error is relevant.

Preview results This study shows that [TODO: put result here]

3 Methods

3.1 Model

The speciation model used in this investigation is a constant-rate protracted speciation model.

A protracted speciation model assumes that species have at least two states: a species is either a good or incipient species. A good species is a species recognized as such, where an incipient species is not yet. Both good and incipient species can generate new incipient species, at the speciation-initiation rates b_g and b_i respectively. An incipient species can become a good species at the speciation-completion rate λ. Both good and incipient species can go extinct at rates μ_i and μ_g respectively (see also figure 1).

The simplest member of the protracted speciation famility is the constant-rate protracted speciation model, which assumes that all rates are constant in time.

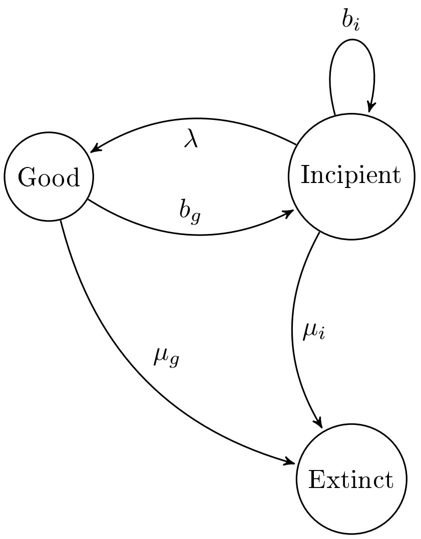

Figure 1: The states and transitions of a species. b_i: speciation-initiation rate of incipient species. b_g: speciation-initiation rate of good species. λ: speciation completion rate. μ_i: extinction rate of incipient species. μ_g: extinction rate of good species. Figure after Etienne et al, 2014, Evolution

There is no relationship known between the speciation rate (often called λ) of the constant-rate birth-death model and a combination of the speciation-initiation λ and speciation-completion rates of the protracted birth-death model (b_g and b_i). When setting $\lambda \rightarrow \infty$, the model used falls back to a constant-rate (non-protracted) birth-death model.

3.2 Workflow

3.2.1 Creating incipient species trees

From a parameter combination, a 'true' protracted constant-rate birth-death incipient species tree is simulated in the R programming language [10], using the PBD package [2].

Speciation-completion rate The speciation-completion rate λ is pivotal for this study, as when $\lambda \rightarrow \infty$ the model falls back to a birth-death model and satisfies the instantaneous speciation assumption of the tools used.

Speciation initiation rate, extinction rate and crown age The parameter values for the speciation-initiation rates b, extinction rate μ and phylogeny crown age t_c need to be balanced, as $t_c b \gg \mu$ results in overly taxon-rich phylogenies, where $t_c \mu \gg b$ results in extinction of all lineages. As a starting point, we used the parameters used by [4]. For simplicity, we assume species can give rise to new species, independent of species status, thus $b_i = b_g$. This assumes that speciation is caused mostly by vicariance and the accumulation of independent mutations. Additionally, we assume a species can go extinct independent of species status, so$\mu_i = \mu_g$.

3.2.2 Creating species trees

A species tree is a phylogeny with one lineage per species. Assuming instantaneous speciation, every lineage represents a species per definition. For protracted speciation one needs to define which subspecies is chosen to represent a species. One such way is to randomly sample one lineage per species. For the sampling being representative, this approach would have to be repeated multiple times, which is computationally infeasable. Instead, we chose (non-randomly) the two most extreme species trees: of each species, we pick the lineages that has been present longest (the oldest) and shortest (the youngest)

3.2.3 Creating a DNA alignment

From each species tree, n_a DNA alignments were simulated following a Jukes-Cantor nucleotide subsititution model using the phangorn R package [13].

In creating DNA sequence alignments from the phylognies, a mutation rate r and DNA sequence length l_a need to be balanced as well. As a starting point, we chose the DNA sequence length to be 1kb, as this a common gene sequence length, but we also included one and two orders of magnitude bigger. The mutation rate is chosen to match a mutation rate as in nature [TODO: find that value].

3.2.4 Creating a BEAST2 posterior

From these DNA alignments, a posterior containing phylogenies and parameter combinations were constructed using the BEAST2 software package [1]. Per alignment, n_b BEAST2 runs were performed, each with an MCMC length of l_m, to see if both runs result in similar posteriors.

The BEAST2 priors used were as follows:

For the site model, the default parameters were used: which is the Jukes-Cantor substitution model.

For the clock model, the default parameters are used, which is a strict clock prior, with a clock rate of 1.0.

For the tree prior, the 'Birth Death Model' was selected and its parameters kept at the default uniform birth-rate (range 0-10^5, initial value 1) and default uniform birth-rate (range 0-1, initial value 0.5).

3.2.5 All parameters used

Table 1 shows all parameter values used.

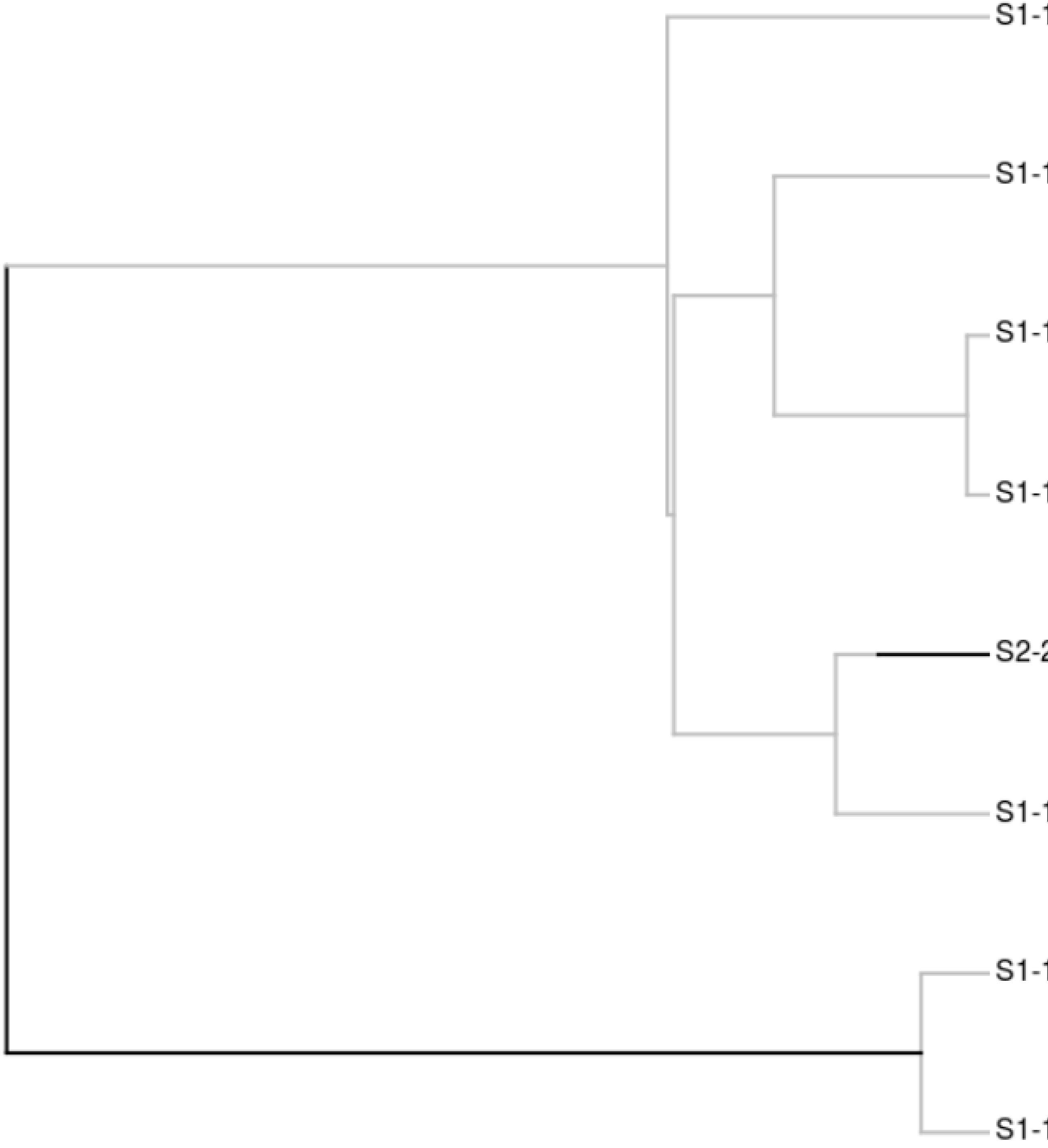

Figure 2: Example incipient species tree and its two species trees. 'youngest' denotes that the most recent lineage is chosen to represent a species, 'oldest' selects the lineage that has been present longest. Because the simulation algorithm, the oldest lineage has the lowest subspecies index (the third number), where the youngest lineage has the highest subspecies index.

Parameters	Values
$b = b_g = b_i$	0.1, 0.5, 1.0
λ	0.1, 0.3, 1.0, 10^6
$\mu = \mu_g = \mu_i$	0.0, 0.1, 0.2, 0.4
t_c	15
r	10^{-1},10^{-2},10^{-3}
l_a	10^3,10^4,10^5
n_a	2
n_b	2
l_m	10^6

Table 1: Parameters used

3.3 Analyzing the results

These posteriors were analyzed using bash (`www.gnu.org/software/bash`) scripts and the R programming language [10], using the packages rBEAST [8] to process BEAST2 output files, ape [9] and ggplot2 [14] for plotting and testit [15] for debugging. All the scripts can be downloaded from `https://github.com/richelbilderbeek/Cer2016`.

How well do two BEAST runs repeat (from the same alignment)? Very good

How similar are the results of different alignments (of the same species tree)? Good

How similar are the results of two different species trees (from the same gene tree)? Similar

The effect of sequence length and mutation rate The effects of sequence length and mutation rate are ...

Number of taxa The number of taxa ...

Difference in nLTT plots Protracted speciation

Histogram of errors Histogram of errors

4 Results

5 Discussion

The protracted speciation model creates trees with less taxa, A constant-rate protracted birth-death model is used, this research could easily be modified to compare other protracted birth-death models, for example, a diversity-dependent protracted birth-death model. There has been now work done on the diversity-dependent protracted birth-deatg model yet.

This research assumes that speciation is allopatric, because in the simulation of the DNA alignments, there is no gene flow anymore between two taxa of still-the-same species.

References

[1] Remco Bouckaert, Joseph Heled, Denise Kühnert, Tim Vaughan, Chieh-Hsi Wu, Dong Xie, Marc A Suchard, Andrew Rambaut, and Alexei J Drummond. Beast 2: a software platform for bayesian evolutionary analysis. *PLoS Comput Biol*, 10(4):e1003537, 2014.

[2] Rampal S. Etienne. *PBD: Protracted Birth-Death Model of Diversification*, 2015. R package version 1.1.

[3] Rampal S Etienne, Bart Haegeman, Tanja Stadler, Tracy Aze, Paul N Pearson, Andy Purvis, and Albert B Phillimore. Diversity-dependence brings molecular phylogenies closer to agreement with the fossil record. *Proceedings of the Royal Society of London B: Biological Sciences*, page rspb20111439, 2011.

[4] Rampal S Etienne, Hélène Morlon, and Amaury Lambert. Estimating the duration of speciation from phylogenies. *Evolution*, 68(8):2430–2440, 2014.

[5] Rampal S Etienne and James Rosindell. Prolonging the past counteracts the pull of the present: protracted speciation can explain observed slowdowns in diversification. *Systematic Biology*, 61(2):204–213, 2012.

[6] Richard G FitzJohn, Wayne P Maddison, and Sarah P Otto. Estimating trait-dependent speciation and extinction rates from incompletely resolved phylogenies. *Systematic biology*, 58(6):595–611, 2009.

[7] Sean Nee, Robert M May, and Paul H Harvey. The reconstructed evolutionary process. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 344(1309):305–311, 1994.

[8] olli0601. *rBEAST*.

[9] E. Paradis, J. Claude, and K. Strimmer. APE: analyses of phylogenetics and evolution in R language. *Bioinformatics*, 20:289–290, 2004.

[10] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2013.

[11] Daniel L Rabosky and Irby J Lovette. Explosive evolutionary radiations: decreasing speciation or increasing extinction through time? *Evolution*, 62(8):1866–1875, 2008.

[12] James Rosindell, Stephen J Cornell, Stephen P Hubbell, and Rampal S Etienne. Protracted speciation revitalizes the neutral theory of biodiversity. *Ecology Letters*, 13(6):716–727, 2010.

[13] K.P. Schliep. phangorn: phylogenetic analysis in r. *Bioinformatics*, 27(4):592–593, 2011.

[14] Hadley Wickham. *ggplot2: elegant graphics for data analysis.* Springer New York, 2009.

[15] Yihui Xie. *testit: A Simple Package for Testing R Packages*, 2014. R package version 0.4.