BIOINF702 Semester 1 2019, Assignment 3
Sebastian Dunn, sdun067, 4610092

Assignment 3 Discussion

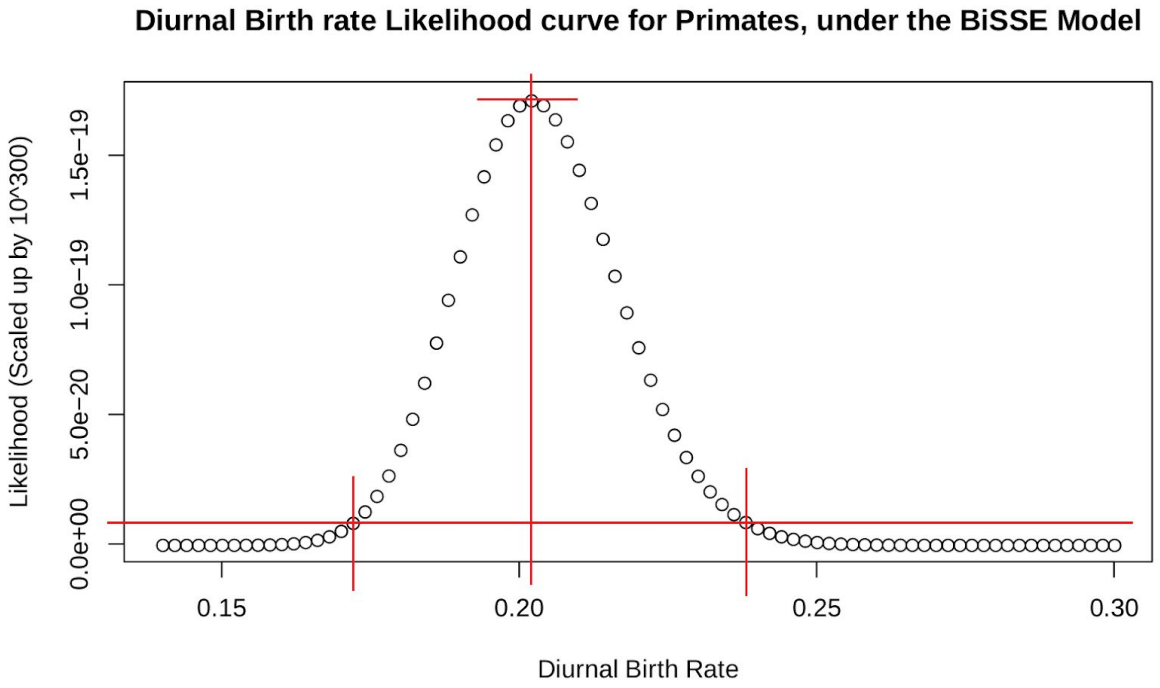
Dataset

My dataset is a tree with 233 tips, representing 233 out of 367 known extant primate species. For convenience, I treated the dataset as complete. This tree is actually a super tree, constructed by analysing different source trees that came from different data types. At the time of publication (2006) it was the most complete tree of the Primates order available.

Divergence times were estimated and calibrated using overlapping fossil sequence data and applying a molecular clock. If there wasn't any age data for a node, it was estimated using a pure birth model.

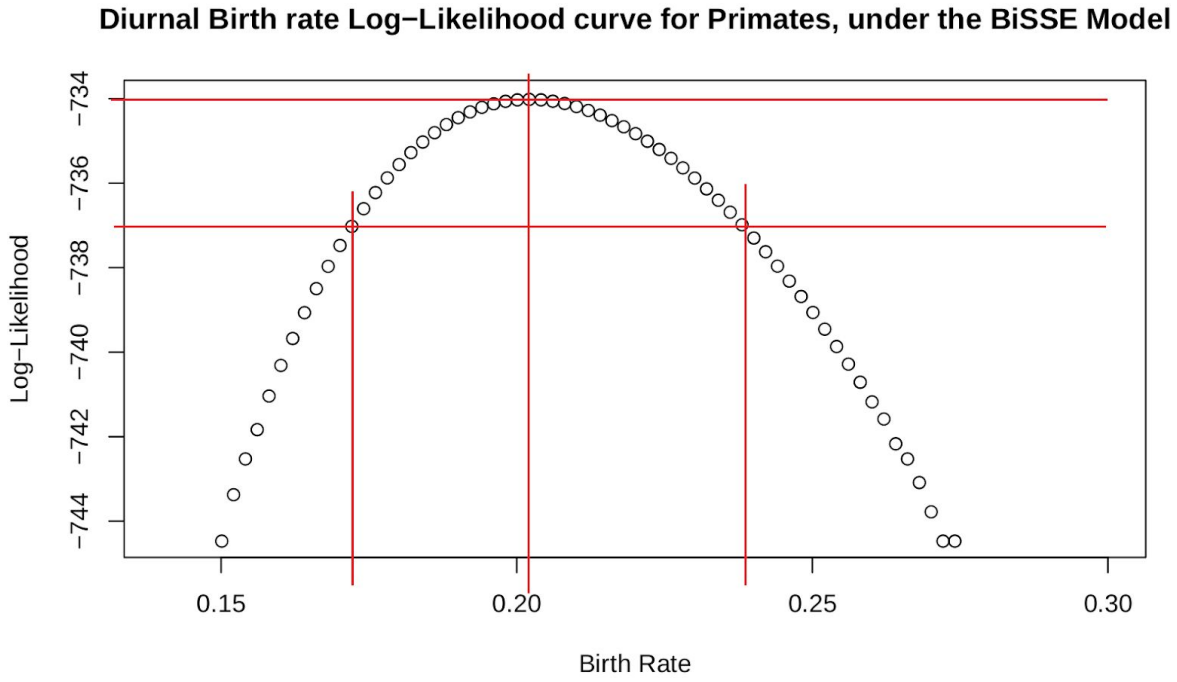
The dataset came from the PhD dissertation of Rutger Vos in 2000, published 2006, called "Inferring Large Phylogenies: The Big Tree Problem", where constructing the super tree was the main push of the dissertation. I accessed it through the RevBayes BiSSE tutorial page: https://revbayes.github.io/tutorials/sse/bisse.html . This version of the dataset has had polytomies resolved using the Kuhn et al (2011) method, and was re-published in 2012 by Magnuson-Ford and Otto.

Interpreting the plots



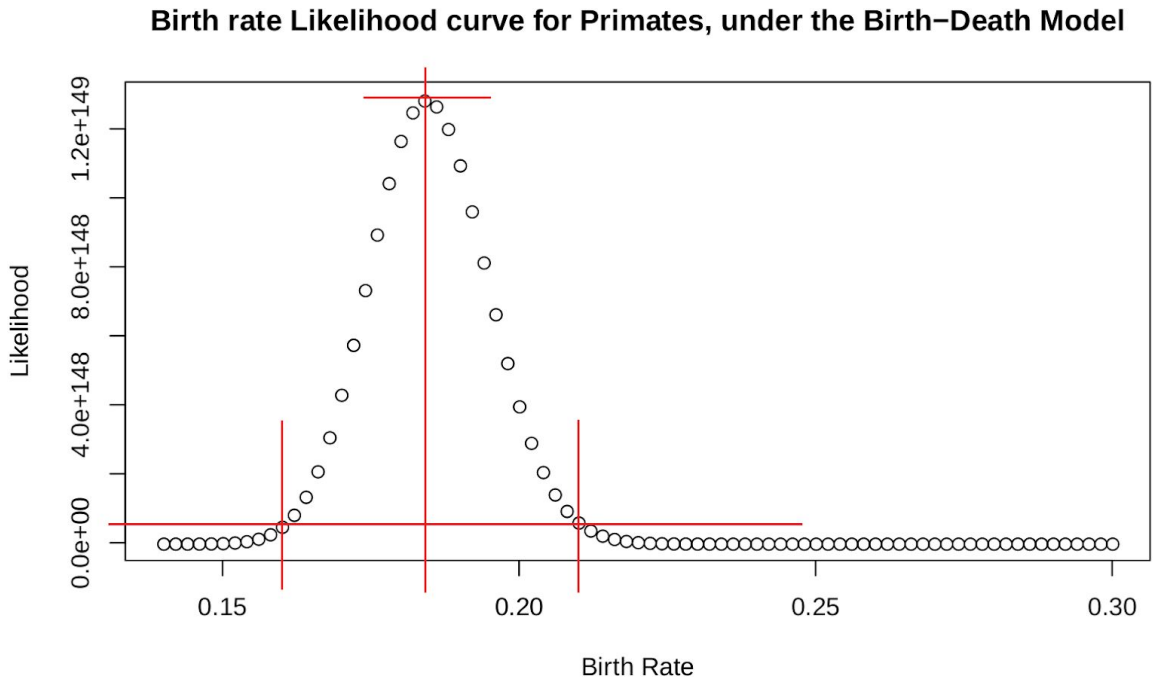
BiSSE Likelihood curve:

- Max Likelihood is around 1.7 x e^{-319} at around 0.202 birth rate.
- 95% confidence interval is approximately 0.17 - 0.24 for the birth rate.



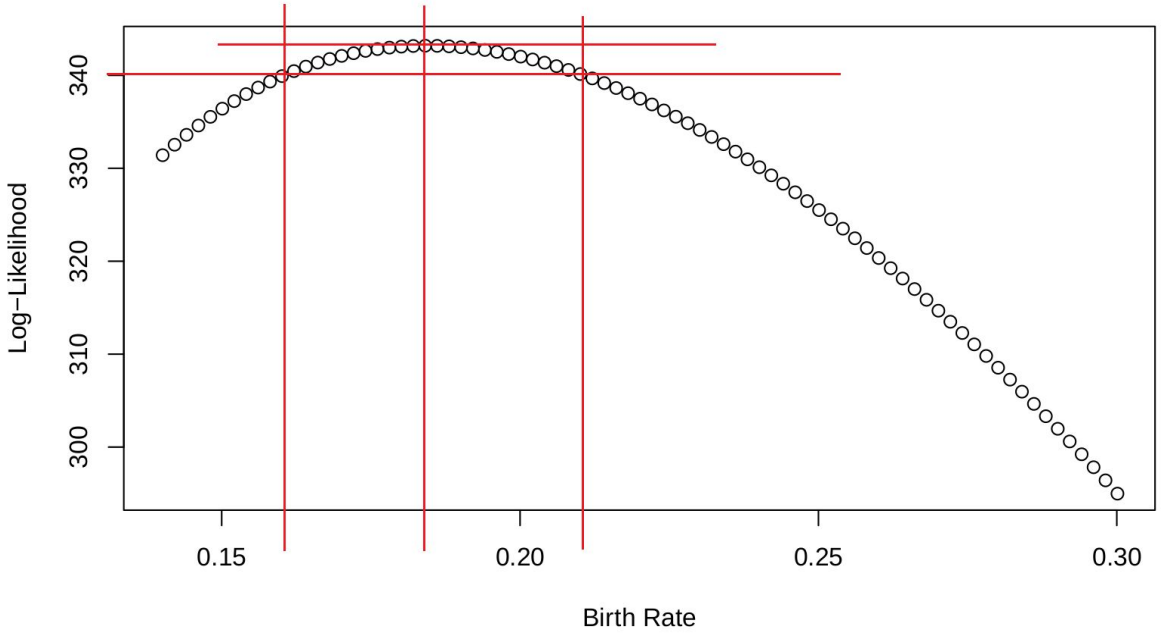
BiSSE Log-Likelihood curve

- Max log-likelihood is around -734 at around birth rate 0.202.
- Approximate 95% confidence interval is from about 0.17 - 0.24 again, which is what we expect.



Birth-Death Likelihood curve

- Max likelihood is around 1.27 x e^{149}, with a birth rate around 0.185.
- Approximate 95% confidence interval is around 0.16 - 0.208 in birth rate.

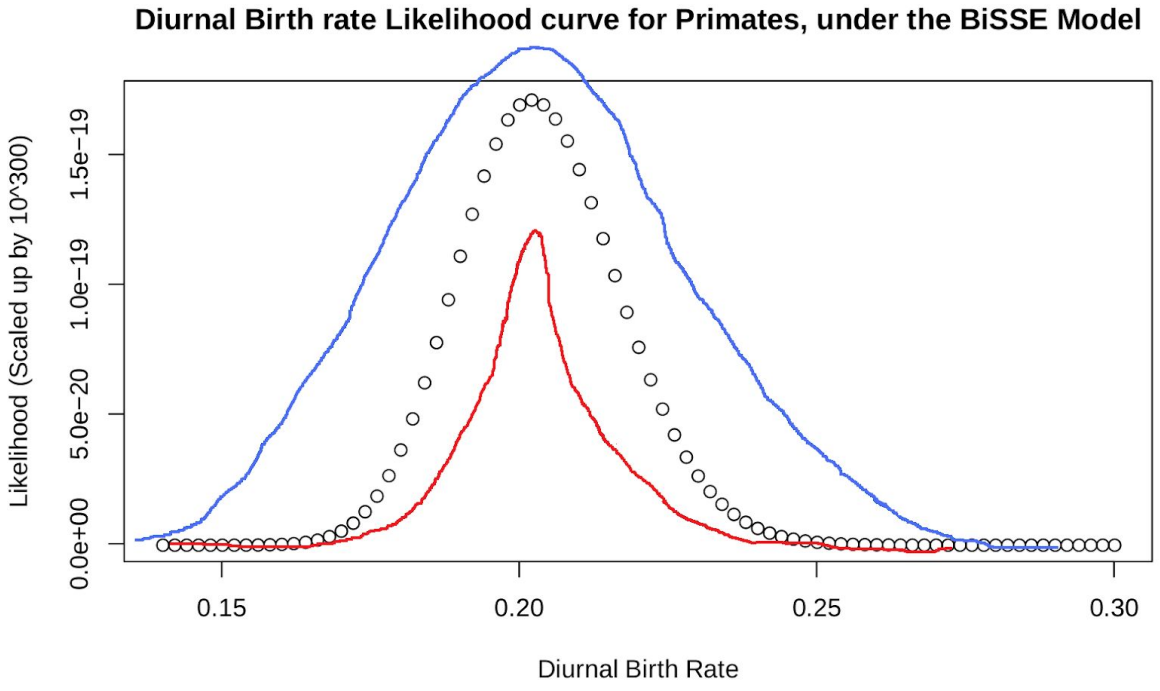


Birth-Death Log-Likelihood curve

- Max lnL is around 343, with a birth rate of around 0.185.
- Approximate 95% confidence interval is around 0.16 - 0.208 in birth rate.

Varying sample size

- Varying the sample size of our tree will change the likelihood plots produced.
- With more samples, overall likelihoods values are going to decrease, but the sharpness of the likelihood curve will increase with a linked decrease in confidence interval range.
- Conversely, with fewer samples, we have higher likelihood values but a less steep curve and a wider 95% confidence interval.
- See below: red curve is increased sample size, blue curve is decreased sample size.



Lessons, surprises and challenges

I've taken three main lessons from doing this assignment: 1) a lot more experience in coding in R and what is available in the language; 2) understanding likelihood calculations a lot more; and 3) how model/likelihood/parameter comparisons are done in practice.

Up until this point my work with statistical models has been mostly theoretic, but working through this assignment made me bridge the gap between theory and practice, and also showed me that the practice is very accessible.

I learned a lot about R and what is available - I got quite familiar with the ape Phylo object, and learned how to simulate tree data under pretty much any model. This let me play around with sample sizes and generating rates, which drove home the lesson that even if you know the rates you generated data with, the best estimated parameters can still be quite different.

I also became a lot more comfortable working between the Likelihood and Log-Likelihood space, doing all of the direct comparisons between the two for the same set of data.

The most surprising part of this assignment was actually discovering what is available in R. I knew there was a large ecosystem, but I hadn't realised how accessible these complex statistical functions are. Freely being able to create simulated data, test models, do parameter maximisation, solve series of ODEs, and pretty much anything else I can think of - these are the advantages of R I have been looking for.

The most challenging part of this assignment was coding the BiSSE likelihood function. It forced me to learn a lot of R syntax and tricks for moving, manipulating and naming data. But the major challenge was understanding deSolve's ordinary differential equation solver framework. Getting my head around the format for the inputs and outputs was difficult, differing between those that vary at each timestep and those that don't, but I am glad I did it and think I have a good result.

To overcome this challenge I read around, looked at examples, and most importantly tried things! R being an interpreted language, there is definite advantage in being able to try out things on the fly to see how they work. I even got into some real debugging.

My BiSSE likelihoods do not exactly match those that Densitree produces. There are a lot of steps in the process where we can vary, but two that I know we do differently are:

- Treatment of the extinction rates down the branches (I feel like I've implemented mine differently to other examples I saw, but I'm not 100% sure how)
- The number of timesteps allocated to each branch. I used a naive 200 steps on each branch, but I know Densitree scales this number based on the branch lengths.

References

Vos, R (2000). Inferring Large Phylogenies: The Big Tree Problem. PhD Thesis. Universiteit van Amsterdam. Available at http://www.sfu.ca/~amooers/papers/Vos_Dissertation06_chV.pdf (Accessed: 10 June 2019)

Kuhn T.S., Mooers A.Ø., Thomas G.H. 2011. A Simple Polytomy Resolver for Dated Phylogenies. Methods in Ecology and Evolution. 2:427–436. 10.1111/j.2041-210X.2011.00103.x

Magnuson-Ford K., Otto S.P. 2012. Linking the Investigations of Character Evolution and Species Diversification. The American Naturalist. 180:225–245.10.1086/666649