Replication of the Gekkotan Phylogenetic Tree

Elizabeth Glynne B.

18 April 2019

**Introduction**

A big component of science lies in repeatability. One of the things discussed in class is how the various parameters of the analyses used can influence the resulting trees, either in branch lengths or resulting topologies. Though, as discussed, when the proper methods are used, often the differing methods will result in comparable answers.

As an evolutionary morphologist, I rely on molecular phylogenies created by other scientists to run my analyses regarding morphological character evolution. I do not create phylogenies based on morphology since the selection of characters can be subjective and far fewer in numbers compared to the sequences used. Therefore, it is important to me to be able to understand how the molecular phylogeny is created and why certain parameters were selected as compared to others.

There are two main molecular phylogenies that are currently used for geckos: one created on a broad scale incorporating as many taxa within Squamata as possible (Pyron et al., 2013) and another that uses just Gekkotan phylogeny with a limited number of species (Gamble et al.,2015). Dr. Pyron’s work has come under scrutiny for missing data in his analyses, which is why individuals who work with geckos will gravitate towards the use of Dr. Gamble’s phylogeny. Therefore, for my future work with morphological data, I will be using the phylogeny from Gamble et al.’s 2015 publication, particularly for evaluating character state evolution. As I will be relying on this phylogeny for future studies, I want to replicate their analysis. I will be looking specifically into if the use of different tree-building methods would have affected the resulting topology or branch lengths. I want to replicate their Bayesian analysis and also perform analyses using parsimony and maximum likelihood methods. Given that I will be using the same data and the same or comparable parameters, the three different analyses (Bayesian Inference, Maximum likelihood, and Maximum Parsimony) will yield different topological results.

**Methods**

*Data and General Parameters.* The data from the publication was download from the listed dryad repository (Gamble et al. 2015). The data file included the 5 nuclear coding genes (RAG1 (1074 bp), RAG2 (366 bp), C-MOS (284 bp), ACM4 (477 bp), PDC (397 bp)), one mitochondrial fragment (protein coding ND2 gene), and the associated tRNAs (1414 bp)) from the 280 taxa used in the study (including 264 gecko species and 16 outgroups). For each analysis, Gallus (the chicken), was identified as the outgroup.

I relied on the partition schemes found within the published study for the use of this replication attempt. This included a BIC partition scheme as determined from PartitionFinder v1.0.1 (Lanfear et al. 2012). There were three partitions identified, including: all nuclear gene data, first and second codon positions of the mtDNA dataset plus tRNAs, and third codon positions of the mtDNA dataset plus tRNAs. I also used the same model of sequence evolution, GTR+G, as determined from PartitionFinder in the publication.

*Bayesian Inference.* The original analysis was done using BEAST 1.7.4 (Bayesian Evolutionary Analysis Sampling Trees). For the purposes of this study, I used MrBayes, in an attempt to see how the results would compare with the original BEAST analysis (Huelsenbeck, 2001). Using the previously stated partitions, the partitioned analysis ran for one million generations, and printed the frequencies every 2,500 iterations. Upon completion, a stepping stone analysis was ~~then~~ conducted. The stepping stone ran for one million generations and printed the diagnostic frequency every 2,500 iterations. Summaries were then completed for both the parameter and the tree files, with the consensus tree being retained. The consensus tree file was then viewed and evaluated using FigTree v.1.4.4 (Rambaut, 2012).

*Maximum Likelihood.* The maximum likelihood analysis was completed using RAxML-NG (Koslov et al. 2018).

*Parsimony.* The parsimony analysis was completed using TnT (Tree Analysis using New Technology, Giribet 2005). The nexus data file was converted to a .*tnt* file using Mesquite v.3.6 (Madison & Madison, 2018). Prior to the analysis used for evaluation, a preliminary search was conducted to evaluate which of the methods would result in the lowest scoring tree. The thought process was that if the lowest scoring tree was found using those parameters, then the tree space being searched for the analysis was decent for the more exhaustive searches. Using all of the parameters was found to result in the lowest score and were then used for a more extensive search. A new technology search was then conducted using a sectorial search (both random and consensus), the ratchet alternating between equal weights, with the drift and tree-fusing using the default settings. The search began with a random seed of 13, running for over two hours, to find the most parsimonious score ten times. All trees were retained after the search. A consensus tree was built after and saved to the tree file, both with and without branch lengths.

**Results**

*Bayesian Inference.* In running the original analysis the estimated mean marginal likelihood of the partitioned model was **X**.In running the stepping-stone sampling, the marginal likelihood of the partitioned model was estimated to be **X**. When looking at the files in Tracer, the marginal densities of the two runs combined leveled off at 302,900.

* Discuss ss, tracer results and tree
* Discuss family level topology
* Discuss any genus level discrepancies
* Compare to Gamble et al directly
* Conclude on overall similarity topologically

*Maximum Likelihood.*

The maximum likelihood score was …. Assessing reliability: bootstrapping

* Discuss bootstrapping
* Discuss family level topology
* Discuss any genus level discrepancies
* Compare to Gamble et al directly
* Conclude on overall similarity topologically

*Parsimony.*

In the analysis there were 41 trees found with the maximum parsimony score of 803,339. A consensus tree was completed upon running the bootstrap analysis with the same new technology search parameters as listed above.

* Discuss family level topology
* Discuss any genus level discrepancies
* Compare to Gamble et al directly
* Conclude on overall similarity topologically

**Discussion**

*General.*

Summarize each of the analysis compared to Gamble et al.

Note any oddities noticed in the analysis and discuss possible explanations.

Discuss any thoughts on possible parameter setting limitations of this study.

*Comparison of methods.*

>Things to answer:

Were there problems observed with maximum parsimony? LBR

Were there problems observed with ML?

Were there problems observed with Bayes?

Were there differences in ML compared to MP?

Were there differences between ML and Bayes?

Were there differences between Bayes and MP?

Concluding statement.

*References*

Gamble T, Greenbaum E, Jackman TR, and Bauer AM. (2015). Data from: Into the light: diurnality has evolved multiple times in geckos. Dryad Digital Repository. doi:10.5061/dryad.97b50.

Gamble, T., Greenbaum, E., Jackman, T. R., & Bauer, A. M. (2015). Into the light: diurnality has evolved multiple times in geckos. Biological Journal of the Linnean Society, 115(4), 896-910.

Gamble T, Greenbaum E, Jackman TR, Bauer AM (2015) Data from: Into the light: diurnality has evolved multiple times in geckos. Dryad Digital Repository. https://doi.org/10.5061/dryad.97b50

Giribet, G. (2005). TNT: tree analysis using new technology.

Huelsenbeck, J. P., and Ronquist, F. (2001). MRBAYES: Bayesian inference of phylogenetic trees. Bioinformatics, 17(8), 754-755.

Kozlov, A. M., Darriba, D., Flouri, T., Morel, B., and Stamatakis A. (2018). RAxML-NG: A fast, scalable, and user-friendly tool for maximum likelihood phylogenetic inference. bioRxiv. doi:10.1101/447110

Lanfear, R., Calcott, B., Ho, S. Y., & Guindon, S. (2012). PartitionFinder: combined selection of partitioning schemes and substitution models for phylogenetic analyses. Molecular biology and evolution, 29(6), 1695-1701.

Rambaut, A. (2012). FigTree v1. 4.

Maddison, W. P. & D.R. Maddison. (2018). Mesquite: a modular system for evolutionary analysis. Version 3.6 <http://mesquiteproject.org>

*Figures*

Parsimony Tree

Figure 1. Consensus tree from the maximum parsimony analysis. Each color coordinates to a family within the infraorder Gekkota as seen in the legend.

Likelihood Tree

Figure 2. Consensus tree from the maximum likelihood analysis. Each color coordinates to a family within the infraorder Gekkota as seen in the legend.

Bayesian Tree

Figure 3. Tree from the Bayesian inference analysis. Each color coordinates to a family within the infraorder Gekkota as seen in the legend.

Bayesian Tree

Likelihood Tree

Parsimony Tree

Figure 4. The resulting analysis trees collapsed to major groups for easy comparison. Colored by family.