Molecular Phylogenetics

Jordan Lillard

April 28, 2019

**Introduction**

With the earliest known turtle species dating back to over 220 million years ago, they are among the oldest reptiles in the world. Their history is ancient, even more so than both snakes and crocodiles. The purpose of this analysis is to observe how a species with an extensive history, turtles, have evolved through the ages.

Turtles belong to the reptile order *Testudines.* Since they are the only living anapsid amniotes, their evolutionary relationship is essential in determining the basal condition and early transformations of many amniotic characters (Shaffer, 1997). Recent molecular analyses have showed some discrepancies with the previously established morphological hypotheses which led Crawford to perform a genome-scale analysis of turtle phylogeny (Crawford, 2014).

In the research below, three different models - Maximum Parsimony, Distance Analysis, and Bayesian Analysis - are utilized to analyze the entire genome to construct a single resolved phylogeny.

**Materials & Methods**

*Data & Parameters*

The data was acquired from a Crawford’s 2014 study. The data includes 32 turtle/ingroup taxonomic units (OTUs) and six outgroups (three lepidosaurs [Sphenodon and two squamates], two archosaurs [a crocodilian and a bird], and one mammal [human]; Table 1) for analysis. The 32 ingroup OTUs represent all of the major lineages of turtles: Pleurodira, Trionychia, Testudinoidea, Chelonioidea, Chelydridae, and Kinosternoidea. The data also includes samples from all 14 traditionally accepted families (Turtle Taxonomy Working Group, 2014). Crawford created a database of UCE loci and prepared FASTA files for alignment. Using MAFFT, a monolithic FASTA file was aligned. The resulting alignments were trimmed using phyluce. This data was split in partitions based on their substitution model, and a concatenated dataset was produced for analysis. I used this concatenated dataset, with the partitions and parameters defined by Crawford’s study.

*Maximum Parsimony*

I performed maximum parsimony analysis using PAUP\*. 100 replications were done using a heuristic search to find the maximum parsimony tree. All trees were saved and, the 50% Majority Rule Consensus tree was found and reconstructed with FigTree (Figure 1). The tree was rooted with the mammals following the approach of a recent analysis that confirms the archosaurs affinities of Testudines (Field, 2014).

*Distance Analysis*

Using PAUP\* again, pairwise distances were used to calculate the distance matrix. The distance matrix was calculated from the nucleotide sequences, under Jukes-Cantor model of sequence evolution. This was then used to estimate phylogenetic trees with the neighbor-joining clustering algorithm and the UPGMA clustering algorithm. The resulting trees were reconstructed with FigTree (Figure 2 and Figure 3).

*Maximum Likelihood*

I referred to the paper’s maximum likelihood analysis of the data. RAxML version 8.0 with the “GTRGAMMA” option and 10,000 bootstrap replicates (Figure 4).

*Bayesian analysis*

I performed Bayesian analysis using MrBayes version 3.2.7a for 500,000 iterations (4 chains; burn-in 25%; thinning: 500. The resulting tree was reconstructed with FigTree (Figure 5).

**Results**

*Maximum Parsimony*

In the analysis, a total of 780,308 rearrangements were tried and one tree was retained with a score of 74,749. Because only one tree was retained, I decided to rerun the heuristic search with 1,000 replicates. The same tree was retained with a maximum parsimony score of 74,749. The bootstrap method with the heuristic search was then done with both 100 and 1,000 replicate and yielded the same tree both times (Figure 1). This tree is consistent with the Maximum Likelihood tree from Crawford’s paper (Figure 4).

*Distance Analysis*

UPGMA failed to create an accurate tree with multiple taxa located in incorrect families. The phylogenetic tree predicted by neighbor-joining was rooted with mammals (Field, 2014). The resulting tree was almost identical to our reference tree (Figure 4), correctly identifying monophyletic Pleurodira and Cryptodira. It also identified Trionychia as sister to all other taxa (Durocryptodira). Within Durocryptodira, Testudinoidea is not found to be sister to all other taxa (Americhelydia). Instead, Kinosternoidea is found to be sister to all other taxa. It fails to correctly identify sister taxa Testudinoidea and Americhelydia.

*Bayesian Analysis*

The recovered phylogeny (Figure 5) is identical to that recovered from the maximum likelihood approach. The data confirms a monophyletic Pleurodira and Cryptodira. The traditional families (Chelidae, Pelomedusida, and Podocnemididae) were recovered with Pleurodira. Within Cryptodira, Trionychia is the sister lineage to all other taxa (Durocryptodira). Durocryptodira contains sister taxa Testudinoidea and Americhelydia.

**Discussion**

*Distance Analysis*

While all five trees do generally show the same phylogenetic relationships, they are not exactly the same. Maximum parsimony, maximum likelihood, and bayesian analysis yielded the same phylogenetic hypotheses. The distance methods, neighbor-joining and UPGMA, did not show the same trees. It can be inferred that this is due to the assumptions made by those two methods. The two clustering algorithms use distance matrices to predict a phylogenetic tree. UPGMA is only applicable if the sequences evolved according to a molecular clock in which sequences evolved at a constant rate over the whole tree. It produces an ultrametric tree in which the distance from the root to the leaves is the same for every leaf. UPGMA did not reconstruct the correct tree because different regions of sequences vary at different rates and different lineages of the tree may have different rates of mutation. Neighbor joining does not make the assumption of molecular clock and does not require the data to be ultrametric. It instead evaluates each pairwise distance as nodes on a tree, rather than taxa. It modifies the distance matrix for each pair of nodes based on their average divergence from all other nodes. While the neighbor-joining method is fast and can handle large datasets, it reduces sequence information and gives only one possible tree. This one tree could be good, but we have no way of knowing. For this reason, it makes more sense to analyze our data with techniques that also evaluate likelihood, such as bayesian analysis.

*Bayesian Analysis*

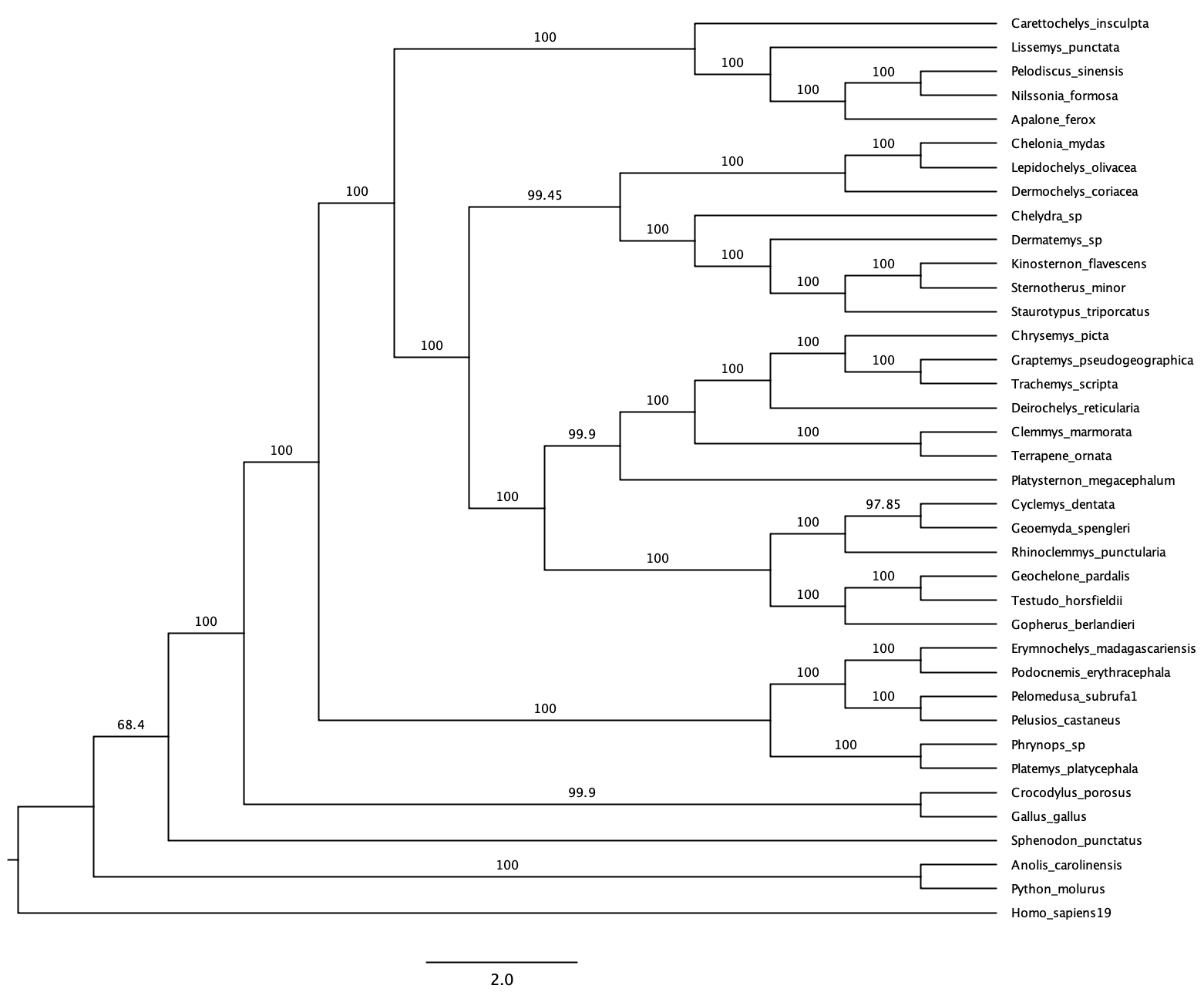
Using bayesian analysis, we found the same phylogenetic tree as that predicted by maximum likelihood. While both analyses depend on likelihood, bayesian analysis also considers prior probability. Maximum likelihood, on the other hand, yields the tree that makes the data the most probable. The data was split into partitions and each partition had its own parameters, due to the fact that the substitution process can differ in different gene regions. Confirmation with TRACER ensured that priors are appropriate and match the expected densities. This allowed us to yield an accurate phylogenetic tree.

*Maximum parsimony*

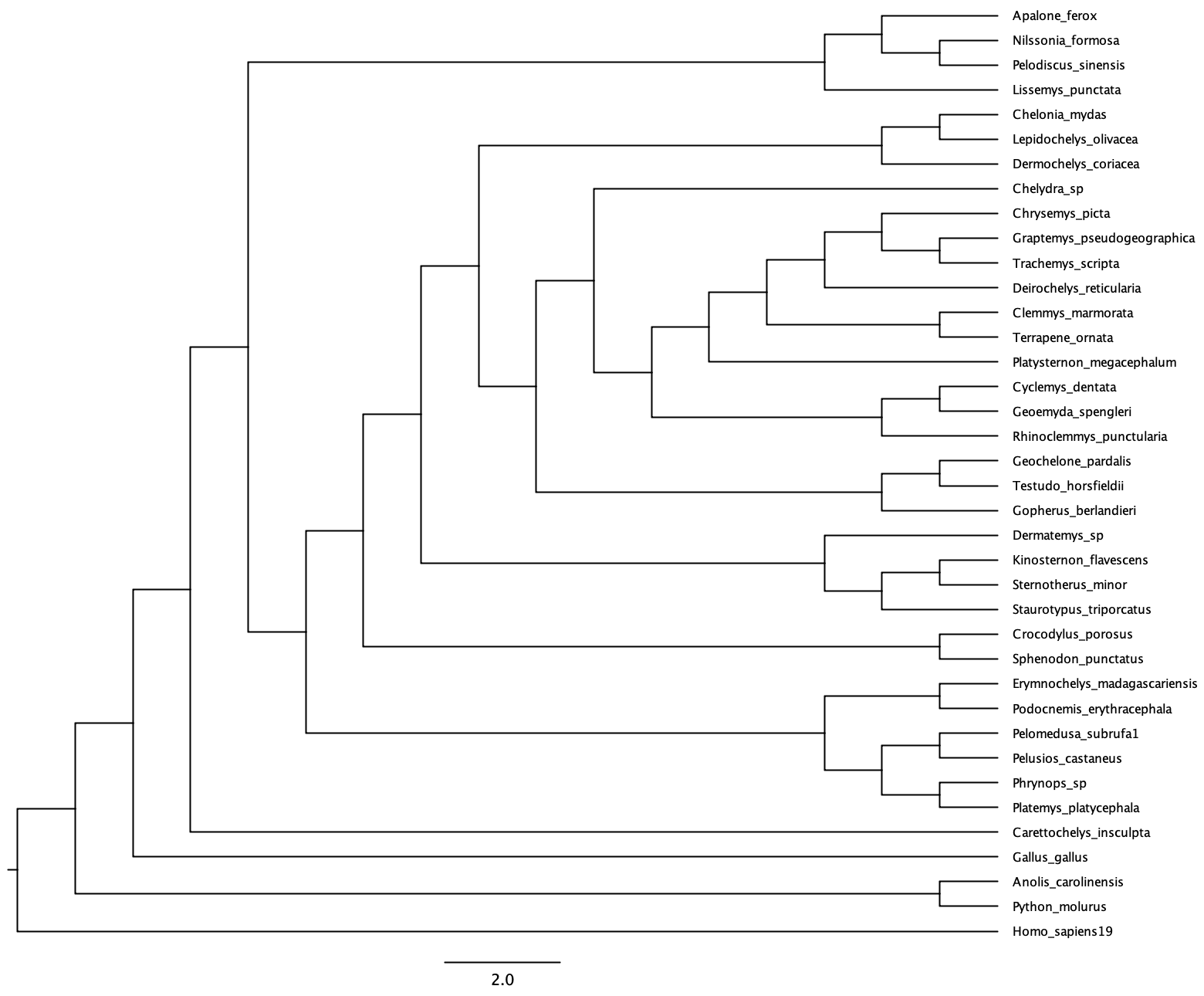
Maximum parsimony predicted the same phylogenetic tree as both our bayesian analysis and maximum likelihood. Maximum parsimony aims to create the phylogenetic tree that minimizes the total number number of character-state changes. It minimizes homoplasy and produces the shortest possible tree. Because of this, it often underestimates the actual evolutionary change that has occured. Another possible weakness of parsimony analysis is long-branch attraction, which occurs when rapidly evolving sequences will have numerous unique mutations. I believe this was avoided due to the fact that we used bootstrapping analysis for confirmation, UCEs for all taxa. This ensured that the sequences were highly conserved, minimizing the changes between even distant taxa.

**Conclusion**

Testudines, or turtles, are one of the most interesting and telling species in regards to amniota evolution. With sequenced genomes and their UCE’s, we were able to construct one consensus phylogenetic tree using maximum parsimony, maximum likelihood, and bayesian analysis. It can be concluded that with this much informative data, more advanced analyses, such as bayesian and maximum likelihood, are more appropriate for inferring phylogenetic trees.



Bootstrap 50% majority rule

Figure 2: Phylogenetic hypothesis based on distance analysis with the UPGMA algorithm.

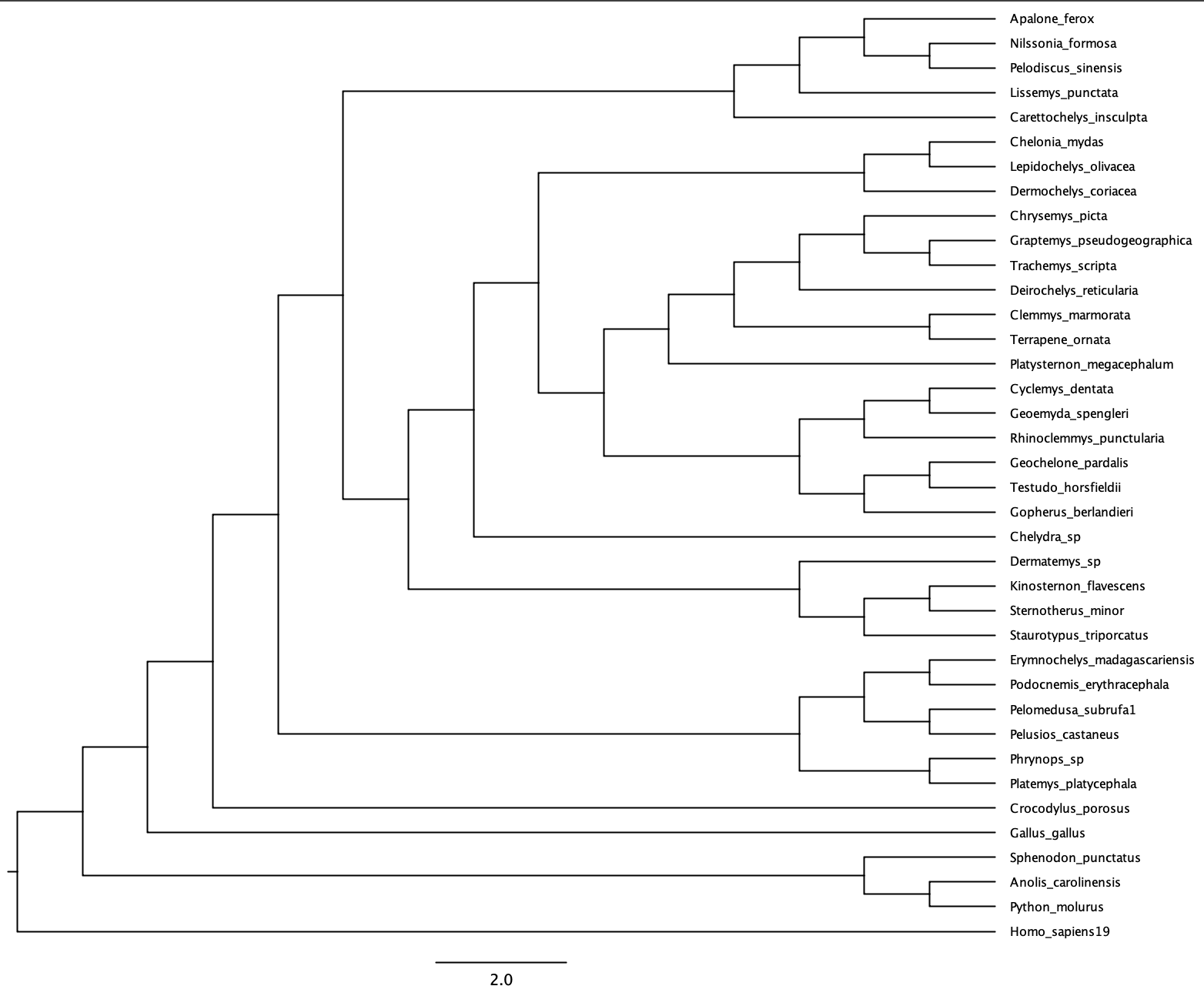


Figure 3: Phylogenetic hypothesis based on distance analysis with the neighbor-joining algorithm.

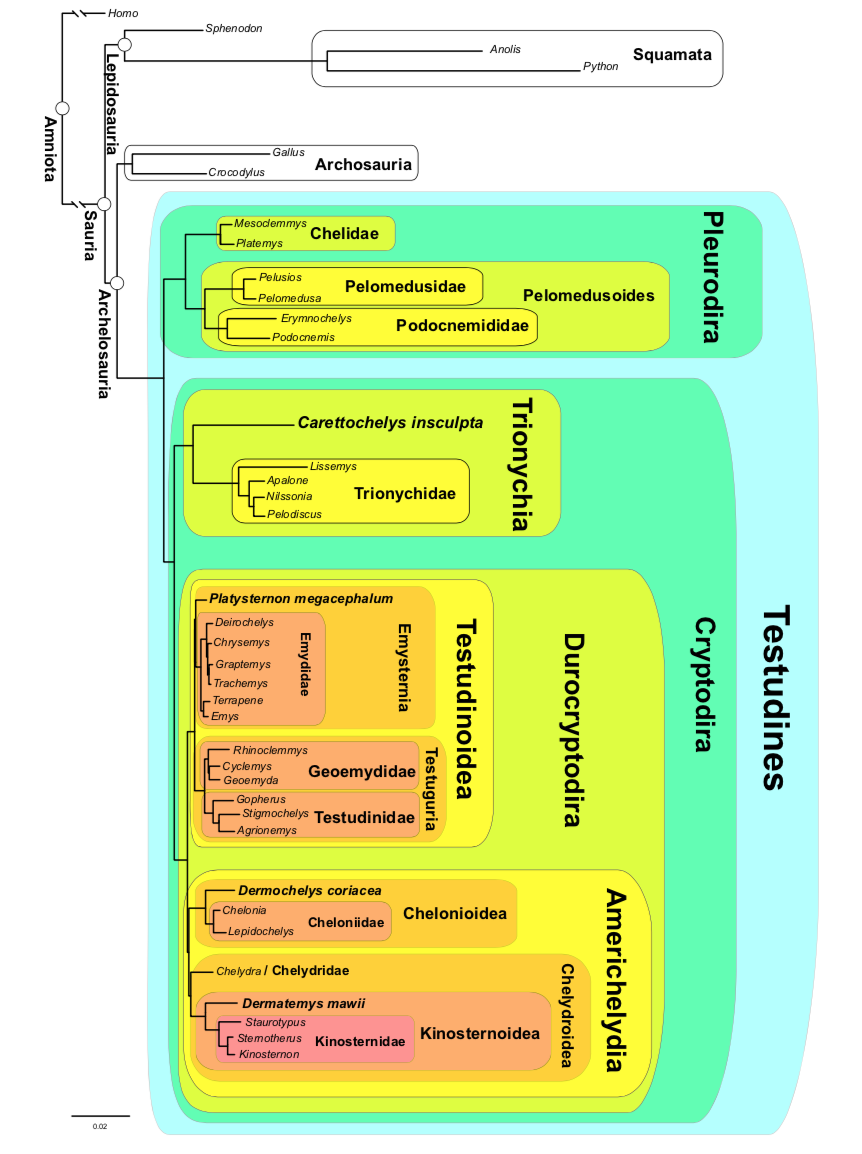


Figure 4. Phylogenetic hypothesis based on RAxML analysis of UCE data showing phylogenetically defined crown clades of turtles (Testudines) (Crawford, 2014).

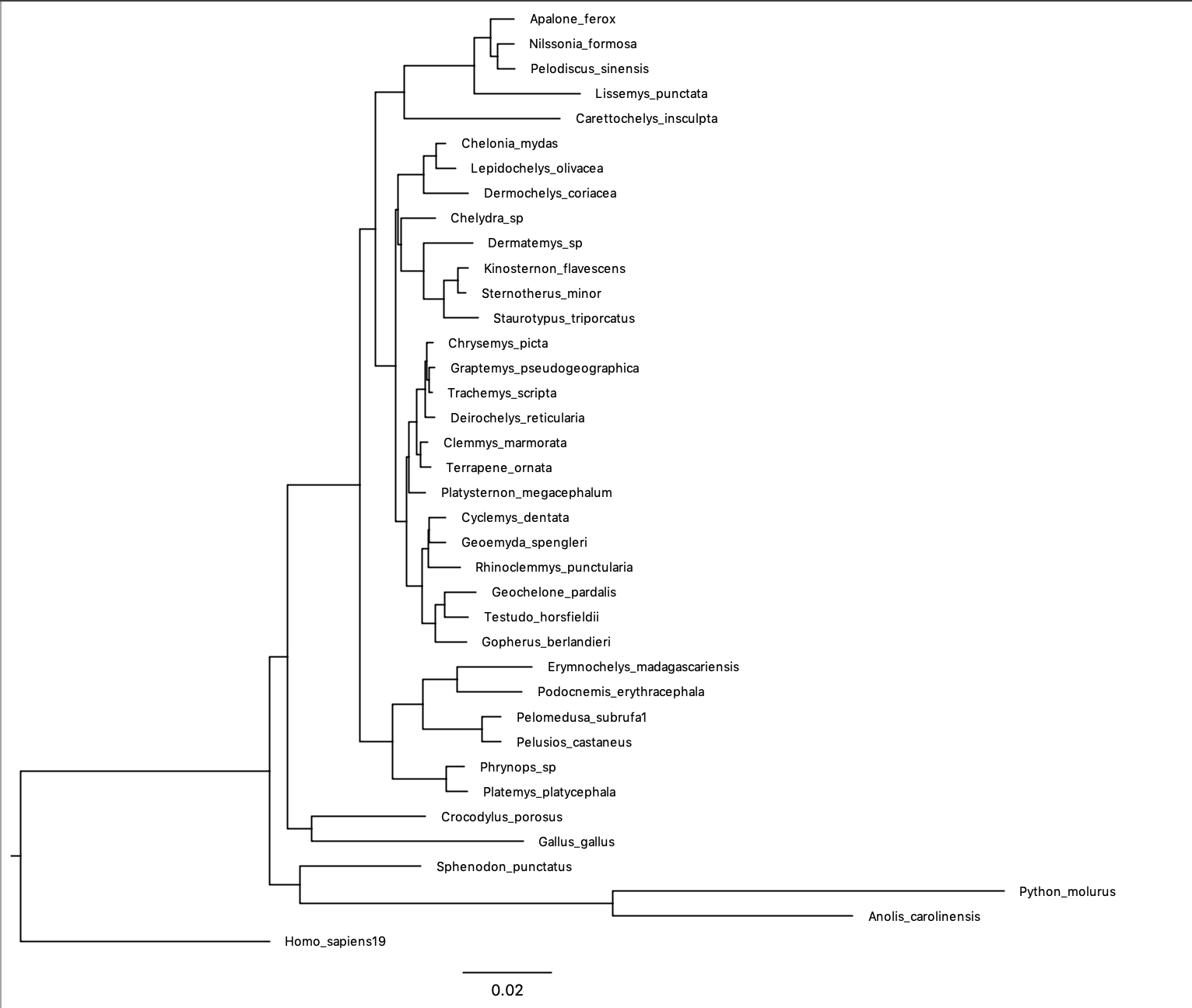


Figure 5: Phylogenetic hypothesis based on MrBayes analysis.

**References**

**Shaffer, H. B., Meylan, P., & Mcknight, M. L. (1997). Tests of Turtle Phylogeny: Molecular, Morphological, and Paleontological Approaches. *Systematic Biology,* *46*(2), 235. doi:10.2307/2413622**

Rambaut A, Drummond AJ, Xie D, Baele G and Suchard MA (2018) Posterior summarisation in Bayesian phylogenetics using Tracer 1.7. *Systematic Biology*. **syy032**. [doi:10.1093/sysbio/syy032](https://doi.org/10.1093/sysbio/syy032)

Turtle Taxonomy Working Group, 2014. Turtles of the world, 7th edition: annotated checklist of taxonomy, synonymy, distribution with maps, and conservation status. Chelonian Res. Monogr. 5, 329–479.

Field, D.J., Gauthier, J.A., King, B.L., Pisani, D., Lyson, T.R., Peterson, K.J., 2014. Toward consilience in reptile phylogeny: miRNAs support an archosaur, not lepidosaur, affinity for turtles. Evol. Dev.. http://dx.doi.org/10.1111/ede.12081.