1 Appendix A: Tree Inference

1.1 Morphological character states

To obtain a realistic value for the probability of having *k* characters states for each simulated morphological character, we randomly selected 100 morphological matrices, each with more than 100 characters each, from TreeBASE (`http://treebase.org/`). We only selected matrices published between 1985 and 2013 and covering 19 taxonomic classes (Chordata, Arthropoda, Annelida, Angiosperm, Gymnosperm and Pteridophyta). This resulted in a total of 22563 characters that had between two and 10 character states. We then extracted the proportion of characters with each number of states (two to 10) to give us an empirical estimate of the average number of character states for each character, as shown in Figure A.1. Most characters have two or three states, therefore we only simulate characters with two or three states, and sample these in proportion to their occurrence in our empirical data (probability of 0.85 for two states characters and 0.15 for three state characters). The code used for this section is available at `https://github.com/TGuillerme/Total_Evidence_Method-Missing_data/blob/master/Analysis/MorphologicalCharacterStates.R`.

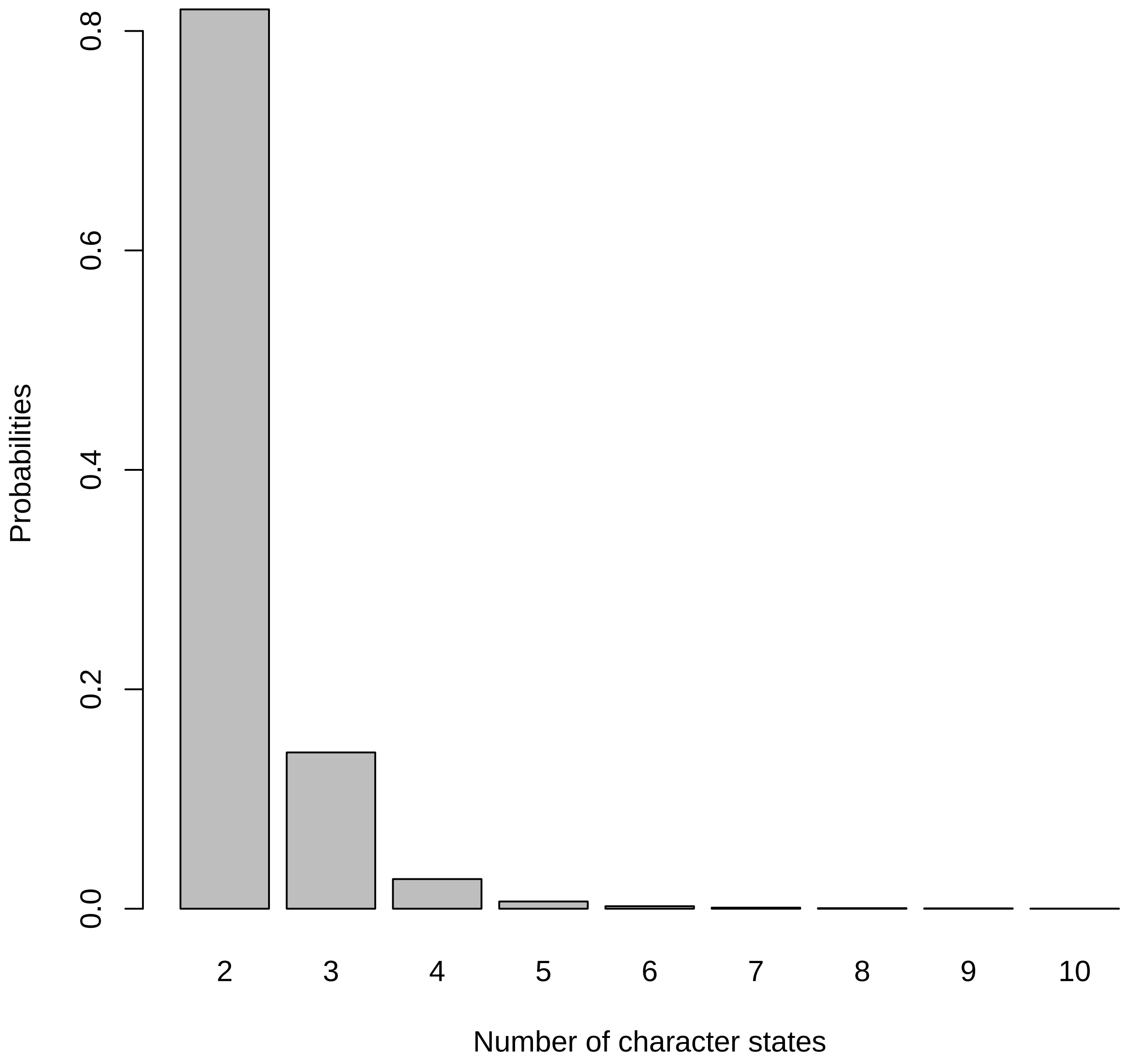

Figure A.1: The proportion of morphological characters with between two and 10 character states extracted from 100 randomly selected empirical matrices downloaded from TreeBASE.

1.2 Tree Inference Software settings

For clarity we have provided the exact settings used in our tree building below.

1.2.1 Maximum Likelihood: RAxML version 8.0.20 Stamatakis (2014)

- Molecular data: GTR + Γ_4 (-m GTRGAMMA)
- Morphological data: M*kv* + Γ_4 (-K MK)
- Support: Rapid Boostrap algorithm (LSR), 1000 replicates

1.2.2 Bayesian: MrBayes version 3.2.1 Ronquist et al. (2012)

- Priors: Molecular data
  - Rates distribution shape (α) = 0.5
  - Transition/Transversion ratio = 2 (β(80,40))
  - Starting tree: "True" tree topology with each branch length = 1
- Priors: Morphological data
  - rates distribution shape (α) = 0.5
- Models
  - Molecular data: HKY + Γ_4
  - Morphological data: M*kv* + Γ_4

- MCMC
  - Two runs
  - Four chains per run
  - Generations $< 5\times10^7$
  - Sample frequency = 1.05×10^4
  - ASDS diagnosis frequency = 5×1^4
  - ASDS < 0.01
  - ESS $>> 200$
  - Burnin = 25%

References

Ronquist, F., M. Teslenko, P. van der Mark, D. L. Ayres, A. Darling, S. Hohna, B. Larget, L. Liu, M. A. Suchard, and J. P. Huelsenbeck. 2012. MrBayes 3.2: efficient Bayesian phylogenetic inference and model choice across a large model space. Systematic Biology 61:539–42.

Stamatakis, A. 2014. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics 30:1312–1313.