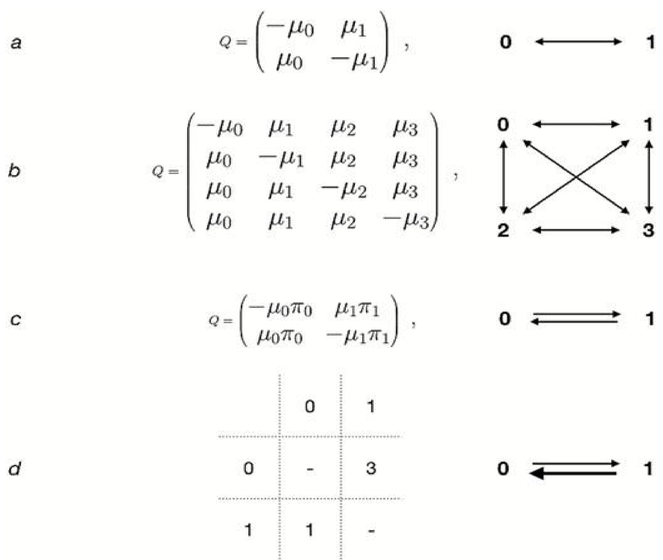
Character State Space Partitioning In Phylogenetic Analysis From Discrete Data

Basanta Khakurel, Tyler D. Tran, Courtney R. Grisby, April M. Wright
Department of Biological Sciences, Southeastern Louisiana University

Introduction

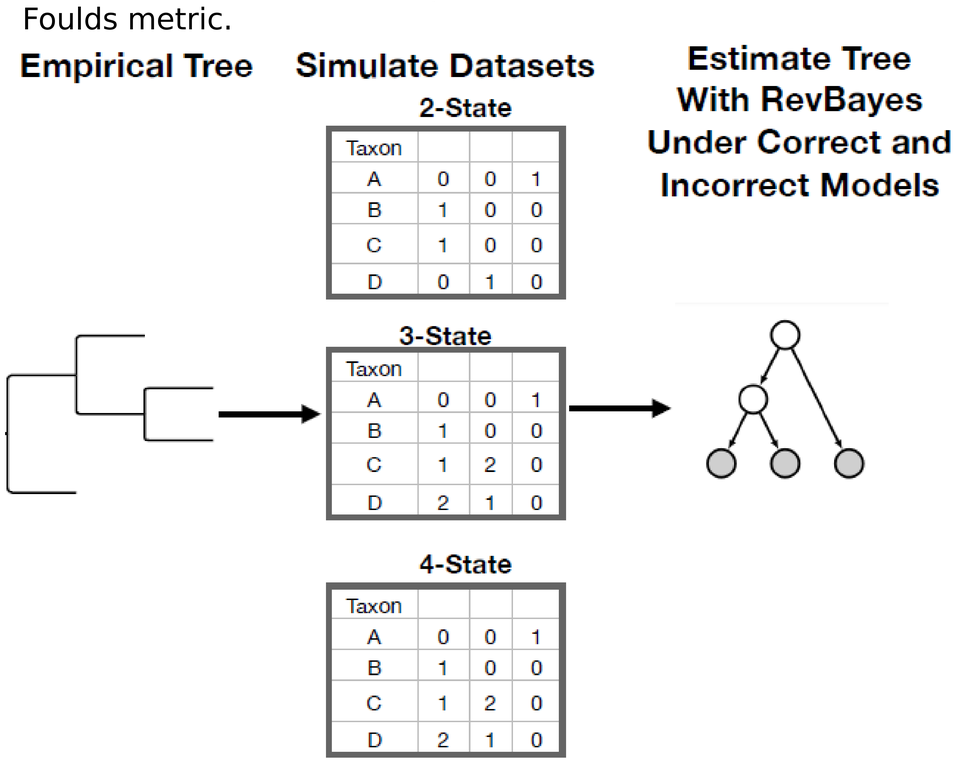
Phylogenetic trees are one of the most useful tools in studying an organism's form and function including its development over time. Phylogenetic trees are generally estimated using Bayesian mathematical model. Molecular data, which is widely used by researchers consist of only four character states - the four nucleotides, whereas the morphological data might have more or fewer character states, including unobserved states. We have looked at the consequences of assuming an incorrect number of character states in a phylogenetic analysis.



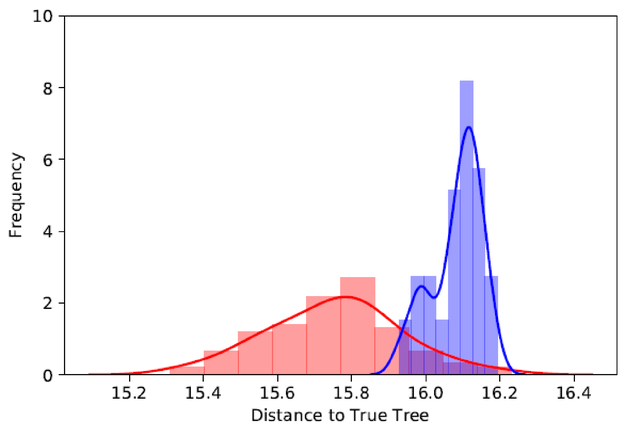
Bayesian methods for discrete characters, express transition probabilities as a Q-matrix. We examined the effects of incorrectly specifying the size of the Q matrix. Figure from Wright (2019).

Methods

We simulated data under the Mk model of morphological evolution in the software RevBayes. Datasets had either 2, 3, or 4 character states. To understand the effect of inappropriate specification of the Q-matrix, we estimated trees from the simulated data under a Mk model with the appropriate number of character states, and under a model with an inappropriate number of character states. For a dataset in which there are 2-state, 3-state, and 4-state characters, the matrix would be partitioned into 3 datasets, corresponding to the three state spaces. Each subset will have its own Q-matrix. In the misspecified model, all the characters were left in a single partition, with the Q-matrix sized according to the character with the largest state number. Trees were then estimated in RevBayes, and their difference to the true tree quantified with the Robinson-Foulds metric.



Results



Distance to the true tree for partitioned (red) and unpartitioned (blue) data.

Conclusion

We used Mk and Bayesian mathematical models to study the effects of incorrect character states estimation in phylogenetic trees. In the above figure, we show that the Robinson-Foulds distance is higher for trees estimated from an unspecified Q-matrix. Assumption of an incorrect number of character states in phylogenetic analyses leads to error in the estimated tree. Because of this error, we have implemented a method in RevBayes, setNumStatesVector(), to automate splitting up a phylogenetic matrix by state number.