Determining Early Life's Evolutionary Rate

A.C. Brown[1], F.U. Battistuzzi[1]
[1]Department of Biological Sciences, Oakland University
Contact: acbrown@oakland.edu; battistu@oakland.edu

Abstract

A chronological framework of life is fundamental to reconstruct the major steps of life's evolution on Earth; from an astrobiological perspective, this can be used to determine the pace of the origin of unique biological processes on Earth and how they relate to the habitability of a planet. However, while the estimation of timetrees in phylogenetic studies has increased due to improved understanding of evolutionary mechanisms, many questions remain on their accuracy. Questions regarding basic assumptions, such as the variation in evolutionary rates among branches of a phylogeny, need to be further investigated to reduce biases derived during the time estimation process. Here we investigate one of the assumptions, the fit of branch-specific evolutionary rates to autocorrelated (AR) and uncorrelated (UR) rate models. Through our large-scale simulation study of prokaryote class and phylum phylogenies, we compare the simulated distributions of ancestor-descendant rate changes to empirical data to determine their fit. We find variable rates among branches but no significant clustering among groups sharing the same common ancestor, even for closely related lineages that would be expected to share similar rates of evolution. Additionally, we find that the empirical data follows neither the AR or UR model, but rather a combination of the two patterns. These results suggested caution when applying these assumptions in divergence time estimations and encourage the use of molecular clock methods that implement fewer assumptions to derive timeline estimations.

Project Overview

Phylogenetic trees are constructed based on comparisons of genetic divergences of sequences among species. The lengths of branches represent the accumulation of substitutions within each lineage; higher numbers of substitutions result in longer branches. Because branches descending from the same ancestor have evolved for equal amounts of time, higher substitution rates within that time correspond to faster evolution. This leads to branch-specific evolutionary rates. Unfortunately, the mode of rate change from one branch to another is highly debated. Two models are currently used, each rooted in phylogenetic or biological properties of empirical datasets (Fig. 1). In a previous study[1], simulated data were used to test the applicability of AR and UR evolutionary rate models in the estimation of parameters for sequences simulated under these models. The two possible scenarios for the estimations were:

Model agreement: the rate model used to estimate parameters is the same used for the simulation of sequences.

Model disagreement: the rate model used to estimate parameters is different from the one used during simulations.

The results showed improved accuracy of estimates when choosing the correct model of rate variation. However, it is currently unknown how each of these models compares to observed rate variation among branches, resulting in difficulty determining the merit of either model. Our aim is to address this question by comparing empirical and simulated rate variation models.

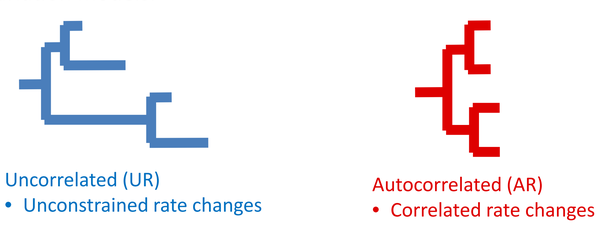


Fig.1 Schematic representation of the two models of rate variation commonly used in phylogenetic analyses

Methods

Simulations were done under 448 sets of parameters[2] modeling each AR and UR for both phylum level and class level data using TvSim[3] and SeqGen[4]. Empirical data and simulations were analyzed with RelTime[5] for relative rates. Percent rate change in rates of ancestor and descendant branches was then estimated ($RC = ((R_a - R_d)/ R_d)$ where RC is the % rate change, R_a is the relative rate of the ancestor and R_d is the relative rate of the descendant) so distributions of rate change could be compared between empirical and simulated AR and UR model data. Statistical analyses are based on the Kolmogorov-Smirnov test run in R[6].

Results

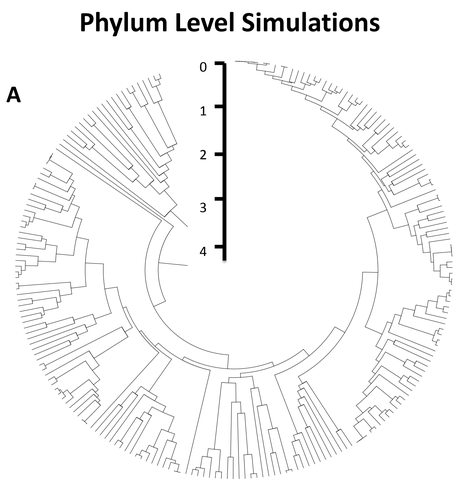
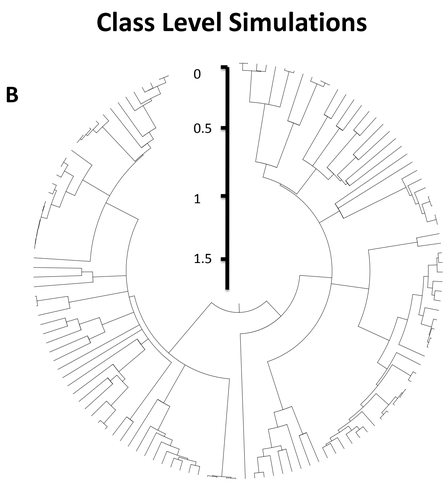


Fig.2 Large datasets were chosen for empirical data. **A:** Phylum level data with 218 species and a root divergence of 4.2 billion years represents a dataset of high variation and older evolutionary history. **B:** Class level data with 129 species and a root divergence of 1.8 billion years was chosen to represents a dataset with more closely related lineages.



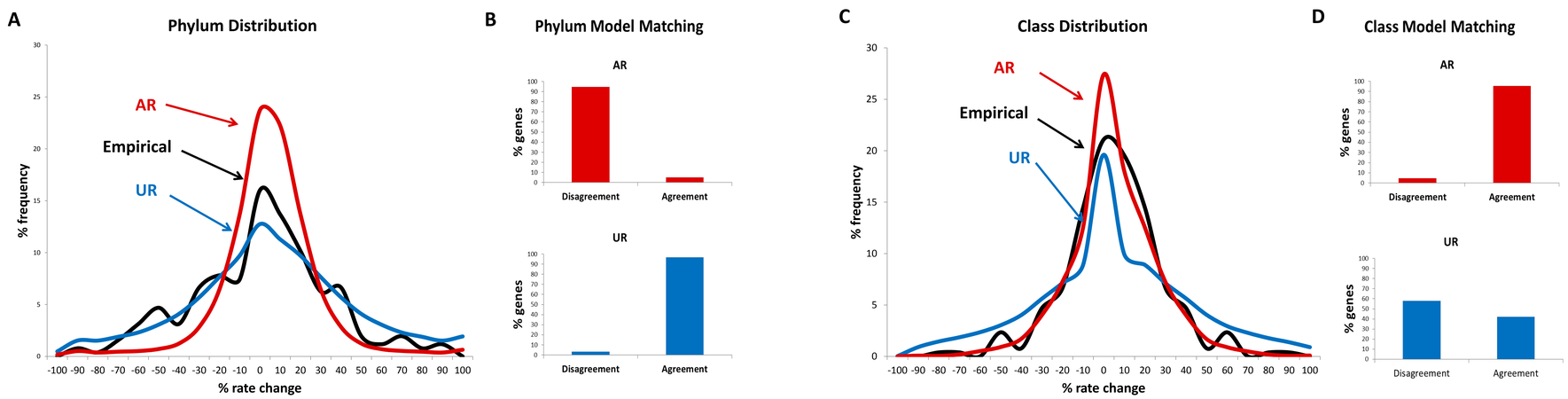


Fig.3 The AR model describes evolutionary changes among closely related species, which are expected to have similar evolutionary rates due to the conservation of biological processes that regulate the accumulation of genetic changes in a genome. The UR model, instead, was proposed to represent phylogenetic trees of distantly related species, which are expected to have highly variable evolutionary rates. **A** and **C:** Distribution of the frequency of rate change between ancestor and descendant lineages for the phylum (**A**) and class (**C**) level phylogeny. **B** and **D:** Statistical tests of model matching reveals model agreement and disagreement. At the phylum level (**B**), the UR model's distribution most closely follows the empirical distribution and at the class level (**D**), the AR model's distribution is closest to the empirical distribution, each with about 95% of the simulated genes in agreement with the empirical genes. At the class level however, the UR model also has similarities in the distribution and about 50% of the simulated genes in agreement with the empirical genes. Ultimately, we find that the models cannot be applied universally and suggest the use of methods deriving timeline estimations under fewer assumptions.

Future Studies

We plan to expand this analysis to include many different phylogenetic levels, from populations to phyla, as well as to phylogenies with varying depths, mammals and plants, to assess the effect of early and recent evolutionary history on model assumptions. This will eventually allow us to assess the accuracy of current phylogenies and timetrees.

Acknowledgments

Thank you to my lab for their support.
This research was supported by Oakland University, Provost award, and the Michigan Space Grant Consortium

References

1. Battistuzzi, F. U., Filipski, A., Hedges, S. B. & Kumar, S. 2010. *Performance of Relaxed-Clock Methods in Estimating Evolutionary Divergence Times and Their Credibility Intervals. Mol. Biol. Evol.* **27,** 1289–1300.
2. Rosenberg, M. S. & Kumar, S. Heterogeneity of Nucleotide Frequencies Among Evolutionary Lineages and Phylogenetic Inference. *Mol Biol Evol* **20,** 610–621 (2003).
3. Kumar, S., Filipski, A., Swarna, V., Walker, A. & Hedges, S. B. Placing confidence limits on the molecular age of the human–chimpanzee divergence. *PNAS* **102,** 18842–18847 (2005).
4. Rambaut, A. & Grass, N. C. Seq-Gen: an application for the Monte Carlo simulation of DNA sequence evolution along phylogenetic trees. *Comput Appl Biosci* **13,** 235–238 (1997).
5. Tamura, K. *et al.* Estimating divergence times in large molecular phylogenies. *PNAS* **109,** 19333–19338 (2012).
6. Ihaka, R. & Gentleman, R. R: A Language for Data Analysis and Graphics. *Journal of Computational and Graphical Statistics* **5,** 299–314 (1996).