

GENOMIC PERSPECTIVES ON THE AMPHIBIAN TREE OF LIFE

Paul Hime, University of Kentucky
2016-2017 Graduate Fellow

EXECUTIVE SUMMARY

The relationships among extant amphibians (frogs, salamanders, and caecilians) have been a longstanding debate in phylogenetics, and previous studies have supported any of the three possible topologies relating these three extant orders of amphibians. This project developed a novel amphibian-specific gene-capture system to target and sequence 220 nuclear genes in a diverse set of 296 amphibian species representing all major amphibian lineages. Using an information-theoretic approach to compare support for the three amphibian orders across the genome, we have discovered substantial discordance across genes in which these three possible models are supported. Phylogenetic information content in many genes in the genome appears to have eroded in the nearly 300 million years since the three amphibian orders diverged. Nonetheless, our results support a model of frogs and salamanders sharing a common ancestor, despite substantial variation in phylogenetic signal across different genes. Our results also suggest a revised hypothesis for the relationships among the extant families of amphibians and support a more recent origin for many of the hyperdiverse lineages of frogs. Overall, this project highlights the power of phylogenomics and a model-based theoretical framework for testing phylogenetic hypotheses in the era of genome-scale evolutionary biology.

RESEARCH CHALLENGE

All organisms trace their ancestry back to a single common ancestor nearly four billion years ago [1]. Yet today, life has diversified into tens of millions of species. Reconstructing these evolutionary relationships is a key aim of the field of phylogenetics, and such insights may inform nearly all aspects of modern biology. The recent advent of genome sequencing technologies has ushered in a new era of phylogenomics in which genetic data generation is no longer a rate-limiting step toward testing phylogenetic hypotheses. Yet, it is becoming increasingly clear that different regions of the genome can support conflicting phylogenetic hypotheses, and reconciling these discordant gene genealogies is a key problem facing evolutionary biologists today. Information-theoretic approaches have great promise for parsing signal from noise in large phylogenomic data sets.

METHODS & CODES

Three possible topologies exist for relationships among frogs, caecilians, and salamanders. The Procera hypothesis supports caecilians+salamanders, the Acauda hypothesis supports frogs+caecilians, and the (canonical) Batrachia hypothesis supports frogs+salamanders. To evaluate support for each of these hypotheses across the genome, we conducted gene-by-gene tests of constrained topology, comparing the maximum likelihood estimates of gene trees for 194 nuclear genes between constraints for the three possible interordinal topologies. Gene trees were estimated in RAxML [2]. We used the Akaike information criterion [3] to quantify the direction and magnitude of support across genes. We then reconciled gene trees into an estimate of the species tree using Astral [4]. This species tree topology was then used in concert with a set of 25 fossil calibrations to estimate divergence times across Amphibia in MCMCTree [5].

RESULTS & IMPACT

This project has demonstrated that there is substantial variation in the amphibian genome for which of the three possible topologies relating the three amphibian orders is supported across genes. Significant numbers of genes are found to support each of the three competing topologies, suggesting that either there has been an erosion of phylogenetic signal over deep time and/or that population-level processes (large effective population sizes and rapid divergences) have led to incomplete lineage sorting and gene tree-species tree discordance. Results from our AIC-based topology testing support the Batrachia hypothesis (frogs and salamanders are each other's closest relatives) at the level of the species tree and suggest that large numbers of genes may need to be analyzed in order to overcome stochastic phylogenetic noise at deep timescales. Our divergence time analyses also provide a revised timescale for amphibian diversification through time and indicate that a rapid increase in net rates of species diversification occurred at the Cretaceous-Tertiary boundary. Analyzing hundreds of genes for hundreds of species has traditionally been computationally intractable for empirical data sets, and this study is one of the first to use an information-theoretic framework to address not only the direction of support for phylogenetic hypotheses across the genome but also the magnitude of that support.

WHY BLUE WATERS

Access to the Blue Waters system provided opportunities to leverage large-scale GPU processors to accelerate the tens of thousands of gene tree estimation analyses needed to implement information-theoretic topology testing. Although many of the analyses I intended to execute on Blue Waters were delayed due to unforeseen issues with my data set, the experiences gained from interacting with NCSA staff and through the Blue Waters Symposia have helped to expand my HPC knowledge and will be invaluable assets in other computational aspects of my planned career.

Paul Hime received his PhD in evolutionary biology from the University of Kentucky in August 2017. He worked under the supervision of Dr. David Weisrock.