Abhijit Chinchani
Diana Lin

Connection Assignment

Field Review: Sequence evolution and phylogenetics
Current Paper: A Fast Likelihood Method to Reconstruct and Visualize Ancestral Scenarios

"A central issue in biology is to recover and understand the evolutionary history of biological entities" [1]. The relationship between the evolutionary events of these biological entities is depicted graphically using phylogenetic trees. The phylogenetic trees are typically inferred using DNA/protein sequences. These evolutionary trees have applications in diverse fields, including ecology, molecular biology, and physiology [2]. All these studies use ancestral character reconstruction (ACR) to trace the origin and evolution of the character of interest. The character state of tips of the tree are usually known, ACR is used to infer the character of the inner nodes of the tree. This global reconstruction over the whole tree describes the evolutionary history of the character and is commonly called an "ancestral scenario" [1].

The fastest and simplest ACR methods are parsimony-based. However, due to their oversimplified assumptions of evolutionary processes, parsimony has limited accuracy [3] and performs worse compared to simple maximum likelihood (ML)-based models [4]. ML and Bayesian approaches are based on probabilistic models of character evolution, which account for evolutionary times and branch lengths. ML-based approaches like Markov chain Monte Carlo (MCMC) procedure are frequently used, where the posterior distribution of ancestor characteristics, tree topography, and model parameters are jointly inferred. However, this comes at a high computational cost and the joint inference cannot be achieved for large data sets.

Currently in the field, state-of-the-art methods either return a unique state for every tree node, where the uncertainty of the inferences is not evident, or return probabilities for every state for each node which makes the ancestral scenario hard to interpret. PastML, a fast likelihood method to reconstruct and visualize ancestral scenarios [1], is a method that overcomes such challenges with its use of a variant of ML, the Marginal Posterior Probabilities Approximation (MPPA) method. Using PastML, each node is associated with a set of most likely states. When there is low uncertainty for the node, a unique state is predicted. On the other hand, when there is higher uncertainty for the node, several states are predicted. Within these several states, thresholding is applied using the Brier score, where lower likelihoods are discarded from the set. This method allows users to see the uncertainty present in certain nodes without misleading assumptions. If a node has multiple states, then those state probabilities are very similar and PastML is unable to make a high confidence prediction of one unique state. However, if a node has one single unique state, the single prediction is one of extremely high confidence. In addition to the reconstruction of ancestral scenarios, PastML also compresses and visualizes the resulting tree on an interactive user interface. In summary, PastML fills a void in the field of phylogenetics by allowing fast and transparent reconstruction of ancestral scenarios, scalable to larger datasets.

References

1. Ishikawa, Sohta A., et al. "A fast likelihood method to reconstruct and visualize ancestral scenarios." Molecular biology and evolution 36.9 (2019): 2069-2085.
2. Soltis, Douglas E., and Pamela S. Soltis. "The role of phylogenetics in comparative genetics." Plant Physiology 132.4 (2003): 1790-1800.
3. Collins, Timothy M., Peter H. Wimberger, and Gavin JP Naylor. "Compositional bias, character-state bias, and character-state reconstruction using parsimony." Systematic Biology 43.4 (1994): 482-496.
4. Gascuel, Olivier, and Mike Steel. "Predicting the ancestral character changes in a tree is typically easier than predicting the root state." Systematic Biology 63.3 (2014): 421-435.