Combining living and fossil taxa into phylogenies: the missing data issue

Thomas Guillerme and Natalie Cooper

Abstract

Living species represent less than 1% of all species that have ever lived. Ignoring fossil taxa may lead to misinterpretation of macroevolutionary patterns and processes such as trends in species richness, biogeographical history or paleoecology. This fact has led to an increasing consensus among scientists that both fossil and living taxa must be included in macroevolutionary studies. One approach, the Total Evidence Method, uses molecular data from living taxa and morphological data from both living and fossil taxa to infer phylogenies with both fossil and living taxa at the tips. Although the Total Evidence Method seems very promising, it requires a lot of data and is therefore likely to suffer from missing data issues which may affect its ability to infer correct phylogenies.

In this study we assess the effect of missing data on tree topologies inferred from total evidence supermatrices. Using simulations we investigate three major factors that directly affect the completeness of the morphological part of the supermatrix: (1) the proportion of living taxa with no morphological data, (2) the amount of missing data in the fossil record and (3) the overall number of morphological characters in the supermatrix. We find that, in a Bayesian framework, difficulties in recovering a stable topology are mainly driven by the missing data in the molecular part of the matrix (for which fossil taxa have no data). In a Maximum Likelihood framework, however, topology is not directly affected by missing data *per se*, but by the number of morphological characters shared among the taxa. Therefore, the two main drivers of incorrect topologies are the overall number of morphological characters and the number of living species with no morphological data.

Our results suggest that, in order to use total evidence methods, one should reduce the missing data in the morphological part of the supermatrix for living species and use a Maximum Likelihood framework to fix the topology prior to the overall Bayesian phylogenetic inference process. We apply our method to a comprehensive data set of both living and fossil primates. We find that using this integrative method modifies previous estimates of rates of body mass evolution within primates.