Empirical distribution mixture models — or how to separate long branches

Dominik Schrempf[1], Nicolas Lartillot[2], and Gergely Szöllősi[1,3]

[1]Department of Biological Physics, Eötvös Lóránd University, Budapest, Hungary
[2]Laboratoire de Biométrie et Biologie Évolutive, Université Lyon 1, CNRS, UMR, Lyon, France
[3]MTA-ELTE "Lendulet" Evolutionary Genomics Research Group, Budapest, Hungary

March 27, 2019

Statistical uncertainties of phylogenetic analyses can be nearly arbitrarily reduced by including more data; an opportunity that is usually readily available given modern sequencing technologies. Especially genome-wide analyses tend to provide high statistical support, for example, in terms of bootstrap values. Statistical support, however, does not measure the accuracy of inferred parameters and phylogenetic trees, but is an indicator of uncertainty in estimates assuming a specific evolutionary model. Systematic errors, for instance due to model mis-specification, are not removed when analyzing more data.

I will explain how ignoring specificity in amino acid usage, which may be a cause of biochemical restrictions, can lead to long branch attraction artifacts. I will introduce phylogenetic models accounting for amino acid specificity and focus on a new empirical method that detects structure in amino acid profiles of high quality alignments. The detected patterns can be used with established, phylogenetic, maximum likelihood software to eliminate systematic errors caused by specificity in amino acid usage in a fraction of the run time that is usually needed.

I will demonstrate the removal of known long branch attraction artifacts from two example analyses; (1) the emergence of microsporidia as sister species of fungi and (2) the emergence of nematodes and flatworms as sister species of arthropodes.