Generalising rate heterogeneity across sites in statistical phylogenetics

Sarah Heaps[1]

[1]*Newcastle University, UK*

Abstract

In phylogenetics, alignments of molecular sequence data for a collection of species are used to learn about their phylogeny – an evolutionary tree which places these species as leaves and ancestors as internal nodes. Sequence evolution on each branch of the tree is generally modelled using a continuous time Markov process, characterised by an instantaneous rate matrix. Early models assumed that the same rate matrix governed substitutions at all sites of the alignment, ignoring the variation in evolutionary constraints. Substantial improvements in phylogenetic inference and model fit were achieved by augmenting these models with a set of multiplicative random effects that allowed different sites to evolve at different rates which scaled the baseline rate matrix. Motivated by this pioneering work, we consider an extension which allows quadratic, rather than linear, site-specific transformations of the baseline rate matrix.

We present properties of the resulting process and show that when combined with a particular class of non-stationary models, we obtain one that allows sequence composition to vary across sites of the alignment, as well as across taxa. Formulating the model in a Bayesian framework, a Markov chain Monte Carlo algorithm is used to explore the posterior distribution. We consider two applications to alignments concerning the tree of life, fitting both stationary and non-stationary models. In each case we compare inferences obtained under our site-specific quadratic transformation, with those under linear and site-homogeneous models.