Gradients do grow on trees: linear-time gradients for high-dimensional phylogenetics

Marc A. Suchard
Departments of Biomathematics, Biostatistics and Human Genetics, UCLA

Abstract

Calculation of the log-likelihood stands as the computational bottleneck for many statistical phylogenetic algorithms. Even worse is its gradient evaluation, often used to target regions of high probability. Order $O(N)$-dimensional gradient calculations based on the standard pruning algorithm require $O(N^2)$ operations where N is the number of sampled molecular sequences. With the advent of high-throughput sequencing, recent phylogenetic studies have analyzed hundreds to thousands of sequences, with an apparent trend towards even larger data sets as a result of advancing technology. Such large-scale analyses challenge phylogenetic reconstruction by requiring inference on larger sets of process parameters to model the increasing data heterogeneity. To make this tractable, we present a linear-time algorithm for $O(N)$-dimensional gradient evaluation and apply it to general continuous-time Markov processes of sequence substitution on a phylogenetic tree without a need to assume either stationarity or reversibility. We apply this approach to learn the branch-specific evolutionary rates of three pathogenic viruses: West Nile virus, Dengue virus and Lassa virus. Our proposed algorithm significantly improves inference efficiency with a 126- to 234-fold increase in maximum-likelihood optimization and a 16- to 33-fold computational performance increase in a Bayesian framework.