*Statistics in Molecular Biology*
IMS Lecture Notes - Monograph Series (1999) Volume 33

# MARKOV CHAIN MONTE CARLO FOR THE BAYESIAN ANALYSIS OF EVOLUTIONARY TREES FROM ALIGNED MOLECULAR SEQUENCES

BY MICHAEL A. NEWTON, BOB MAU AND BRET LARGET[1]

*University of Wisconsin-Madison and Duquesne University*

We show how to quantify the uncertainty in a phylogenetic tree inferred from molecular sequence information. Given a stochastic model of evolution, the Bayesian solution is simply to form a posterior probability distribution over the space of phylogenies. All inferences are derived from this posterior, including tree reconstructions, credible sets of good trees, and conclusions about monophyletic groups, for example. The challenging part is to approximate the posterior, and we do this by constructing a Markov chain having the posterior as its invariant distribution, following the approach of Mau, Newton, and Larget (1998). Our Markov chain Monte Carlo algorithm is based on small but global changes in the phylogeny, and exhibits good mixing properties empirically. We illustrate the methodology on DNA encoding mitochondrial cytochrome oxidase 1 gathered by Hafner *et al.* (1994) for a set of parasites and their hosts.

**1. Introduction.** Stochastic models have long been considered useful for describing variation in the molecular sequences of extant populations (e.g., Jukes and Cantor, 1969; Felsenstein, 1973; Kimura, 1980). Parameters in such models include the phylogeny, which encodes the pattern of evolutionary relationships among populations, and substitution rates, which describe how molecules change over time within populations. It seems quite natural to infer these parameters using the induced likelihood function in some way, but such inference has been difficult in practice because computations can be prohibitively expensive. Owing to the Markovian nature of the standard models, evaluation of the likelihood function follows straightforward recursive equations, and so evaluation is not the difficult part. The difficulty arises with optimization, since the likelihood resides over a complicated parameter space, and seems to admit no simple representation (Felsenstein, 1981, 1983; Goldman, 1990; Yang, Goldman, Friday, 1995). Nevertheless, computer code is available for approximate maximum likelihood calculation (Olson, *et al.*, 1994; Felsenstein, 1995; Swofford, 1996).

Beyond estimation, practitioners have demanded some way to assess uncertainty in aspects of the estimated phylogeny, just as error bars accompany simpler kinds of point estimates. A standard and appealingly simple calculation

[1]B. Larget acknowledges the support of the National Science Foundation.
*AMS* 1991 *subject classifications.* Primary 62F99; secondary 92D20.
*Key words and phrases.* Cospeciation, Metropolis-Hastings algorithm, phylogeny.