The space of labeled trees is big and complicated; for any significant number of leaf nodes and any reasonable probability-based model of evolution it is computationally infeasible to integrate the likelihood function over all trees. Maximum likelihood methods can find optima on this space, but give us little information about the likelihood distribution – useful information for evaluating the validity of the resultant phylogeny and the evolutionary model used. Markov Chain Monte Carlo methods find natural application to this problem. Currently, it is still more current in the literature to estimate confidence using the bootstrap; however, the interpretation of bootstrap results is less clear[2], and since the standard bootstrap usage requires re-running the maximum likelihood procedure many times, it is computationally costly.

There have been a few methods proposed for using MCMC to sample from the posterior of a distribution on phylogenetic trees, e.g. [2]. The methods proposed by Larget & Simon[2], and further explored by Mau et.al.[3], use the Metropolis-Hastings algorithm to construct the appropriate markov chain, and is described for clocklike and non-clocklike trees. MrBayes[?] implements a method for MCMC sampling of the posterior; I don’t know the details.

One of Larget & Simon’s proposal mechanisms on the space of clocklike trees is simply depicted in terms of the topology of the Robinson graph. It is equivalent to the process of, simultaneously for each branch, perturbing the length in a window of size $\delta$, and in the case that one (or more) branches are contracted to zero, picking a new split uniformly out of the possible other consistent splits, and assigning the remaining perturbation length to the corresponding edge. In terms of the polytope representation, it does a random walk in radius $\delta$ (wrt the “taxicab” norm), and upon crossing a boundary, picks one of the other incident faces uniformly at random to enter into. Their other mechanism differs only in constraining the direction the random walker can move in for each step. I intend to look more into this, to see if relatively recent results bounding mixing times using ideas of conductance[4] can be put to use here.

One of the most suprising things for me was to discover the multiplicity of methods used in MCMC techniques to determine “convergence,” and their heuristic nature. Mau, et.al. [3], Larget & Simon[2], and Gray & Atkinson[1] all used time-series methods of analysis: inspecting time-series plots of log-likelihood and autocorrelation to determine appropriate burn-in times and subsampling rates. It was particularly instructive to see the wide range in choices of these values (for instance, subsampling ranged from every 100 trees for a data set of 9 taxa[3], to every 10,000 trees for the set of 87 languages in [1]). The log-likelihood plots shown were not particularly reassuring either — they may have “leveled off”, but were still undergoing seemingly large fluctuations. My skepticism was somewhat allayed by the results (e.g. Table 1 — in all cases shown,
the sampled tree topologies were a tiny fraction of the possible topologies, even across runs from different starting points.

The study of 87 Indo-European languages[1] was particularly impressive, both in the size of their phylogeny, and in the care taken to check assumptions. Re-running the algorithm with different assumptions about data, different rooting techniques, and different priors obtained slightly different topologies, but the consensus trees obtained unanimously supported their main thesis about divergence times. The method used to separate the mutation rate from divergence times, parametric rate smoothing, was interesting, and somewhat suspect in my opinion — it assumes that mutation rates varies somewhat smoothly through branch points.

MCMC sampling seems to be a very useful technique in phylogenetic tree reconstruction. Given the heuristic nature of most methods for determining necessary run lengths and subsampling rates, and doubts about the appropriateness of choosing a prior distribution, it is best to systematically check these assumptions by running multiple chains with different assumptions. Gray & Atkinson’s paper seemed a good example of this. There’s a lot more for me to learn.
Bibliography


