An eTBR proposal for non-binary trees in MCMC Bayesian phylogeny

Blaise Li, Universidade do Algarve, Portugal, blaise.li@normalesup.org
Peter Foster, Natural History Museum London, United Kingdom

MCMC mixing efficiency

Achieving good mixing in Markov Chain Monte Carlo (MCMC) Bayesian phylogenetic analyses is essential to obtain valid posterior probabilities. Convergence of the runs to the region of the maximum a posteriori tree is not a sufficient requirement. Indeed, if the solution space is complex, other regions of the solution space might have non-negligible posterior probabilities. The Markov chains must therefore be able to leave the maximum a posteriori region in order to visit the other regions according to their posterior probabilities. Failure to sample a large enough proportion of the solution space may lead to overestimated posterior probabilities.

Improved convergence

Faced with convergence problems in an analysis using p4 (Foster, 2004), we tested the effects of increasing the boldness of the proposals through the tuning parameters (right column) and through the use of bolder topology modification proposals (bottom row). For each situation, the log-likelihood plot of 10 runs is shown.

With default tunings and LOCAL move as the only topology modification proposal, the runs appear to be stuck in two distinct log-likelihood regions of the solution space. With bolder tunings the runs occupy a wider variety of regions. Runs may start by sampling a low log-likelihood region, but display occasional long jumps which may enable them to reach higher log-likelihood regions.

We expected that the use of the eTBR move would cause each individual run to sample a wider region of the solution space. This does not seem to be the case. A possible reason could be the existence of a sharp contrast in posterior probability between the two main visited regions and the rest of the solution space. The eTBR move however allowed all the runs to converge to the higher log-likelihood region. This indicates that the topology proposals are frequently bold enough to reach that region. The effects of bolder tunings are no longer conspicuous when the eTBR proposal is used.

In analyses of our dataset with a recent version of MrBayes (not shown) using the default proposals resulted in the sampling of a region of very high amplitude of log-likelihood, but lower than what was achieved in the results presented above. Forcing the use of only the LOCAL move resulted in a behaviour similar to what we observed with p4 (runs stuck at various log-likelihood levels, including the same high log-likelihood region as in the p4 analyses).

References

Foster, P. G., 2004. Modeling compositional heterogeneity. Systematic Biology 53 (3), 485–495.
URL http://dx.doi.org/10.1080/10635150490445779

Lakner, C., van der Mark, P., Huelsenbeck, J. P., Larget, B., Ronquist, F., 2008. Efficiency of Markov chain Monte Carlo tree proposals in Bayesian phylogenetics. Systematic Biology 57 (1), 86–103.
URL http://dx.doi.org/10.1080/10635150801886156

Larget, B., Simon, D. L., 1999. Markov chain Monte Carlo algorithms for the Bayesian analysis of phylogenetic trees. Molecular Biology and Evolution 16 (6), 750–759.

Lewis, P. O., Holder, M. T., Holsinger, K. E., 2005. Polytomies and Bayesian phylogenetic inference. Systematic Biology 54 (2), 241–253.
URL http://dx.doi.org/10.1080/10635150590924208

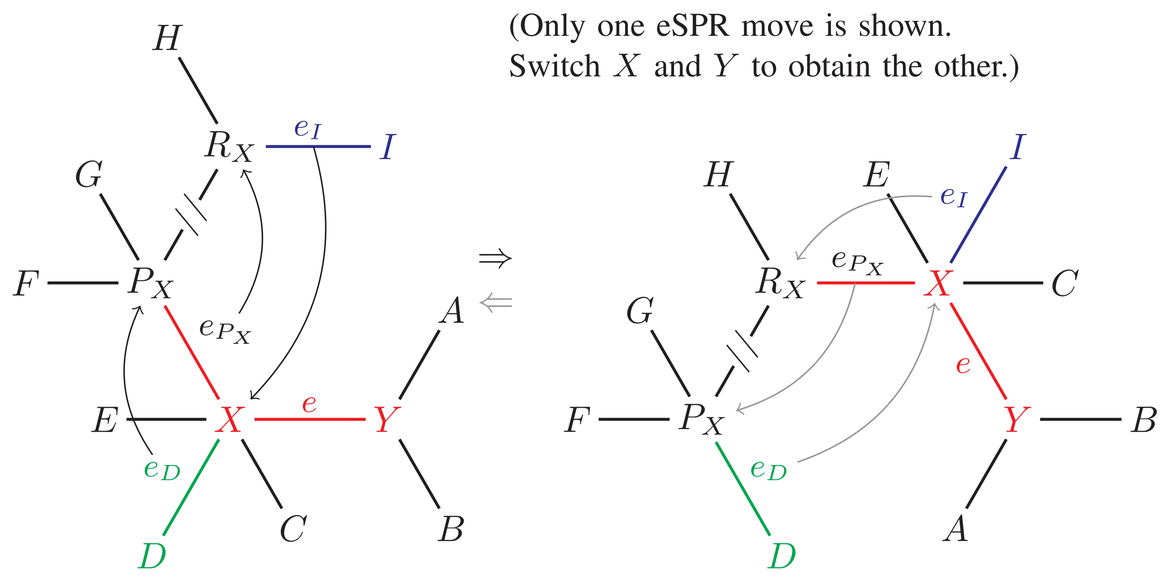
p4 is available at http://code.google.com/p/p4-phylogenetics/.

Bolder topology moves

Some of the topology modification proposals used in MCMC phylogenetic estimation are by design insufficiently bold to allow large jumps in the solution space. If only these types of topology move are used, even the hot chains of a Metroplis Coupled MCMC (MCMCMC) run may become stuck in a limited region of the solution space. This seems to happen when using the LOCAL topology proposal (Larget and Simon, 1999). Lakner et al. (2008) have shown that the extending-TBR move (eTBR) is one of the most efficient topology modification proposals available. In this poster, we present results describing the effects of a quantitative or a qualitative change in the boldness of MCMC proposals, and we describe a modification of the eTBR topology proposal that can accomodate trees with polytomies.

A polytomy-compatible eTBR move

Following the recommendations of Lewis et al. (2005), p4 includes proposals that collapse or create branches. Nodes can therefore have a degree higher than 3. However, the eTBR move described by Lakner et al. (2008) assumes that the tree is fully bifurcating, hence such a move has to be modified in order to be used in p4. We propose the following mechanism:



As the eTBR move described in Lakner et al. (2008), our polytomy-compatible eTBR proposal consists in two extending-SPR moves (eSPR), on both sides of a randomly chosen edge e (in red). The pruning site P_X is chosen at random among the non-Y neighbours of X: this starts the extension process. The extension continues with a certain tunable probability, if there are suitable neighbours available. This determines the regrafting site R_X and one of its neighbours (in blue) between which the pruned edge will be regrafted.

In the original move, X has always 3 neighbours: the other end of edge e (Y); the pruning site (P_X); and a third one, which is detached from X and reattached to P_X. To adapt the move to trees that may be non-binary, one has to decide what to do when X has more than one non-P_X and non-Y neighbour.

We propose to choose one of them at random (in green) to move to P_X, and keep the others attached to X. Choosing to move all such neighbours to P_X would instead lead to an asymmetrical modification of the degrees of the nodes, and the move will not be reversible (and thus not constitute a valid MCMC proposal).

With our current choice, **a reverse move is possible** (in grey).

The boldness of the topological change depends on the number of successful extensions applied to determine the regrafting site.

Conclusions

We confirm that the type of topology move used in MCMC phylogenetic estimation should be considered carefully. Although the use of eTBR did not seem to increase the area sampled by the MCMC runs in our particular test case, it allowed a better convergence of the different runs to a same region of the solution space. If you encounter convergence or mixing problems; check whether the software implements efficient topology modification proposals. Recent versions of MrBayes employ a mix of various proposals (including eTBR by default, but not LOCAL) which seems to be able to achieve an efficient mixing. The causes of the differences of behaviour between MrBayes and p4 MCMC implementations deserve further investigation.

Funding

This work was supported by a Fundação para a Ciência e a Tecnologia (FCT, Portugal) grant to Cymon J. Cox, Centro de Ciencias do Mar (CCMAR) - CIMAR-Lab. Assoc., (PTDC/BIA-BCM/099565/2008).