Section 3 from Castillo-Ramirez et al "Bayesian estimation of species trees: a practical guide to optimal sampling and analysis" chapter in Knowles & Kubatko *Estimating Species Trees*

Some Tips on running the BEST Markov Chain Monte Carlo Algorithm

The most recent version of BEST is found at www.stat.osu.edu/~dkp/BEST where the source code and Windows and Macintosh executables may be downloaded. BEST is also available for use on the BioHPC compute cluster at the Computational Biology Service Unit at Cornell University (http://cbsuapps.tc.cornell.edu/best.aspx).

BEST runs within the Bayesian phylogenetics package MrBayes (Ronquist and Huelsenbeck, 2003) and is thus able to take advantage of the gene-tree estimation infrastructure provided by that program. A BEST analysis will then require the same steps as any Bayesian phylogenetic analysis: i) read the data, ii) set the substitution model(s), iii) set the priors, iv) set the MCMC search rules, v) run the MCMC, vi) check the convergence, and vii) summarize the results. For advice on parts of the program not unique to BEST, it is important to read the MrBayes manual (http://mrbayes.scs.fsu.edu/manual.php). Most of the unique aspects of BEST are specified using the *prset* command in MrBayes that has been extended (steps iii & iv) to allow the user to specify prior probability distributions on the parameters of the gene/species tree model (*ThetaPr* and *GeneMuPr*) and of the MCMC rules (*PoissonMean* and *PropTemp*). In particular:

- BEST uses the inverse gamma prior for the values of effective population sizes of ancestral populations, in the form of the parameter $\theta = 4N\mu$. The inverse gamma prior distribution has two parameters, α (which should larger than 2 to provide a finite variance) and β. The mean of this distribution is $\beta/(\alpha-1)$. The variance is the mean divided by $(\alpha-1)(\alpha-2)$, The distribution is skewed to the right. The parameter theta (Θ) is approximately the number of variable sites per base pair within species expressed on a per site basis. This will differ from locus to locus, so an average across loci might be appropriate to compute (see Edwards and Beerli, 2000). It is recommended to use a fairly low value of α (say $\alpha = 3$) and then fix β so that the mean value is close to the range appropriate for your problem. For example, you might take *Thetapr=invgamma(3,0.004)* if there are about 2 variable sites per pair per thousand sites within species. When only one allele is sampled per extant species, θ is still relevant because it applies to genetic diversity in the ancestral species occurring at each node in the species tree. Although the same prior is used to model θ at all nodes in the species tree, the posterior of θ can differ among nodes.

- BEST allows for different substitution rates across loci with the prior following the uniform distribution with bounds set by the user. The default setting is *GeneMuPr=uniform(0.5, 1.5)* which provides a maximum 3:1 ratio of rates from the slowest to the fastest. When your data has more loci or a wider variation of rates then these bounds should be changed. For example, *GeneMuPr=uniform(0.2, 1.8)* would allow up to a nine-fold variation. Such a prior could be particularly useful if one is combining nuclear and mitochondrial DNA data, whose substitutions rates differ dramatically. It is recommended to keep the center of the distribution at 1.
- BEST takes a set of proposed gene trees from MrBayes and matches it with a species tree in the neighbourhood of the Maximum Tree, a tree that satisfies the constraint that all divergences of species pairs must occur closer to the present than the respective gene divergences occur (Liu et al. 2009). In particular, the Maximum Tree is modified at a random number of nodes following the Poisson distribution, where the mean of this distribution is specified using the *poissonmean* parameter (default value = 5). With larger values a wider variety of species trees will be examined, but this may come at the expense of a smaller acceptance rate.
- BEST applies an annealing step that downweights the influence of the prior at the beginning of the burn-in period in order to move more rapidly into areas of high likelihood. The temperature in the annealing step is cooled at a rate controlled by the parameter *propTemp*. For example, propTemp=0.05 means that the algorithm will use annealing for the first 5% of generations (default value = 0.1). It is not recommended to use the annealing step for longer than half of your intended burn-in period.

Table 4 illustrates the commands used in a BEST analysis of nucleotide sequence data on 4 loci from 22 strains of yeast comprising six species of the genus *Saccharomyces* (Liti, Barton, and Louis, 2006). It also provides a commentary on how the settings might differ from a standard concatenated gene analysis in MrBayes.

It is very important to investigate whether the set of loci being used form an appropriate data set for BEST and how the MCMC settings should be altered from defaults *before* beginning a BEST analysis. As examples,

- Have the loci been investigated for signs of horizontal gene transfer or hybridization? Both of these conditions violate the coalescent model in BEST and such violation may be a source of poor convergence in some data sets (Cranston et al. 2009). BEST should not be used if any HGT, gene flow, or hybridization is present in the data set. BEST could be used cautiously if there is gene flow between sister species, because such gene flow will have the effect of shortening the branch separating those two species, but will otherwise not affect the toplogy. But in general gene flow violates the model in BEST and it should not be used under these circumstances. The effect if removal from a data set of species participating in gene flow is not yet known.
- Do the loci have wildly different substitution rates so that the GeneMuPr prior distributions do not allow posteriors that overlap with each other? If the ratio of the maximum substitution rate to the minimum substitution rate is likely to be larger than allowed by your choice of prior, then the chosen model is not compatible with the data (and the bounds of the uniform prior should be extended). Some researchers have reported challenges when combining nuclear and mitochondrial DNA data in BEST, presumably because of their very different substitution rates. However, further work on this is needed.

- Does one locus appear to dominate the likelihood? This can be checked quickly by finding the likelihood for an estimated gene-tree $\tilde{G}_i$ from that locus on an estimated species tree from the other loci (say $\tilde{S}^{(i)}$, found by a different species tree method such as STAR, STEAC or Deep Coalescence; Maddison and Knowles 2006; Liu et al. 2009). A locus whose likelihood $f(\tilde{G}_i \mid \tilde{S}^{(i)})$ is much lower than the others, or one for which $\left|\log\left[f(\tilde{G}_i \mid \tilde{S}^{(i)})/f(\tilde{G}_i \mid \tilde{S})\right]\right|$ is large, may be dominating the MCMC (here $\tilde{S}$ is an estimated species tree based on all of the loci).

- Do some loci exhibit wildly non-clocklike evolution? BEST employs a fairly cursory method of producing ultrametric gene trees after they are first estimated using the non-clock constrained methods in mrBayes. If loci are strongly violating the molecular clock, branch lengths in the gene trees and species trees could be biased. Because BEST relies heavily on gene tree branch lengths to constrain the search for compatible species trees, violations of the molecular clock could have significant impacts.

Investigating such issues will alleviate most situations where BEST fails to converge due to data sets that do not follow the underlying model assumptions or when the MCMC settings are not appropriate to the problem. For example, Cranston et al. (2009) ran a data set consisting of up to 165 loci from six species of rice, and found that BEST failed to converge after many hundreds of millions of MCMC cycles. However, as these authors discuss, these grass species undergo extensive gene flow, thereby violating the model in BEST. Failures of the MCMC in BEST to converge could often be due to such violations, or to a paucity of data. What passes for sufficient data in a typical concatenated Bayesian analysis may be very insufficient with a species tree analysis. For example, it has been shown that supermatrices with substantial amounts of missing data but nonetheless yield high posterior probability in a Bayesian concatenated analysis can yield very low posterior probabilities in a BEST analysis (Thomson et al. 2008). Although BEST can handle missing loci for some species, we recommend that data matrices should be as complete as possible for BEST to perform accurately. Failure for the MCMC to converge could also be due to inappropriateness of the coalescent prior when dealing with highly diverged data. It may that the coalescent prior behaves poorly when the data are sampled from highly divergent species, even if the coalescent prior is in principle appropriate for those data. Finally, as with any MCMC algorithm, the chain may become stuck on islands of probability, even after hundreds of millions of generations. In these situations, the species tree distribution will typically show good signs of convergence but diagnostics for the individual gene trees appear to fail. Rather than continuing such programs indefinitely, it is better to increase the *nruns* parameter and check if the runs with the highest likelihoods are converging to the same distribution of species trees and whether the other runs are finding islands with high enough posterior probability to force the need for longer chains.

Table 4. Commands and their use in the MrBayes block of a nexus file for running BEST.

MrBayes/BEST command	Comments

begin mrbayes;
outgroup NRRLY969_Unknown;
set autoclose=yes nowarn=yes;
	Species trees are rooted so you must specify an outgroup in the analysis.

charset NEJ1 = 1-1074;
charset EST2 = 1075-3769;
charset HDF1 = 3770-5538;
charset HDF2 = 5539-7428;
partition currentPartition = 4: NEJ1, EST2, HDF1, HDF2;
set partition = currentPartition;
	The 4 loci (NEJ1, EST2, HDF1, and HDF2) are set up as in conducting a concatenated analysis using the partition set capabilities of MrBayes.

taxset species1=1-9;
taxset species2= 10-14;
taxset species3= 15-17;
taxset species4= 18;
taxset species5=19-21;
taxset sspecies6=22;
	The first nine strains are from species 1, strains 10 through 14 are from species 2, etc…
	Note that the taxset command should be used even in data sets with only one taxa per species.

lset applyto=(1) nst=6 rates=gamma;
lset applyto=(2) nst=2 rates=equal;
lset applyto=(3) nst=6 rates=equal;
lset applyto=(4) nst=2 rates=gamma;
	Specify the substitution models for each loci.

prset BEST = 1
Brlenspr=clock:uniform
Thetapr=invgamma(3,0.004)
GeneMuPr=uniform(0.5,1.5)
PoissonMean=5
PropTemp=0.05;
	BEST=1 for species tree, BEST=0 for regular MrBayes. Problems with radically non-clocklike gene trees will fair poorly in BEST. If appropriate, specify clocklike gene trees in MrBayes to help with convergence.
	See the text for advice on uniquely BEST parameters Thetapr, GeneMuPr, PoissonMean, and PropTemp

unlink topology=(all) brlens=(all) genemu=(all) shape=(all) Statefreq=(all) tratio=(all) revmat=(all) ;	*Gene trees are assumed to be independent given the species tree. Thus, unlike a concatenated analysis, it is important to unlink the tree topologies, branch lengths, and other parameters associated with the substitution models for each loci.*

mcmcp
ngen= 20000000
nruns=2
nchains=1
printfreq=1000
samplefreq=1000;
	The extra parameters (population sizes and each gene tree) require much longer chains for BEST. We recommend a minimum of 100 million generations for even small data sets, with larger numbers as the data set size grows. Since gene trees are influenced by the prior from the same species tree in each MCMC cycle, Markov chain coupling is not as profitable in BEST, so nchains of one or two is adequate.

mcmc Stoprule=yes Stopval=0.03 relburnin=yes burninfrac=0.25 samplefreq=1000;	*Run the MCMC as you would for a gene tree analysis*
Execute mydirectory/yeast.nex.sumt;	*BEST v. 2.3 creates a special ".sumt" file with the commands to summarize the results. Edit as necessary. See Thomson, 2009 for advice on displaying the species tree in a program like figtree.*