Laboratory excercise written by Andrew J. Crawford <andrew@dna.ac>
with the support of CIES Fulbright Program and Fulbright Colombia.
`http://filogeografia.dna.ac`
Enjoy!

***Filogeografía**: genética evolutiva espacial.*

BIOL 4211, Universidad de los Andes, Bogotá. Coord.: -074.0657, 04.6012
25 de enero a 01 de abril 2006

Lab 8
MrBayes version 3.1.2.

18 marzo 2006

NB: Always download the latest version of all software
and read carefully the latest accompanying documentation.

The goal for today's lab is to conduct a preliminary Bayesian phylogenetic analysis. You will run a longer, proper analysis again on your own during the coming week. Recall that the goal of Bayesian analysis is **not** to find the best tree but rather to estimate the posterior joint probability distribution of topologies, branch lengths and values of the parameters assumed in our evolutionary model. The posterior probability is proportional the likelihood (aka, probability of the model give the data) multipled by the prior probability. Thus, every parameter (including topology and branch lengths) needs an assumed prior probability, *i.e.*, its probability distribution of possible parameter values *before* collecting the data. For any single parameter of interest, we can obtain the marginal posterior probability by integrating over all other parameters. We can estimate this posterior probability distribution by taking a few thousand samples from a Monte Carlo Markov chain, and then summarizing the results among samples. In Bayesian phylogenetics we are most interested in the set of most probable trees, also known as a "credible set" (again, we are not looking for a single best tree, but a collection of trees that best explain the data) Therefore, in Bayesian phylogenetics we want to create a consensus tree of all our samples taken from the posterior probability distribution. The frequency of any given node or clade (aka, "bipartition") among our sample of trees gives us our estimate of the posterior probability that the node is correct.

Assuming we are not going to collect any more data, we have three issues of special concern in running an adequate Bayesian MCMC phylogenetic analysis.

I. Evolutionary model.
- A) Is our model simple enough that we have sufficient data to support it?
- B) Is our model complex enough to describe the processes that produced the data?
- C) How is our choice of model influencing (biasing?) our estimate of the posterior distribution?

II. Prior distributions on model parameters.
- A) Are they influencing (biasing?) our estimate of the posterior distribution?

III. The MCMC analysis used to estimate the posterior distribution.
- A) Is the MCM chain converging on the true posterior distribution?
  - 1.) Is the MCMC mixing well?
  - 2.) Has it run long enough?
  - 3.) Was the burn-in period long enough?
- B) Are successive samples of the MCM chain independent?

Today we will use *MrBayes* version 3.1.2, which offers some important new features not available in version 3.0. *MrBayes* can be used to analyses amino acid sequences, morphological data and DNA sequence data. Today we will discuss only the latter. *MrBayes* is not the only software for Bayesian phylogenetics. Advanced students should consider using *BayesPhylogenies* (`http://www.rubic.rdg.ac.uk/meade/Mark/`) instead of *MrBayes*. For analysis of RNA gene sequences, one might use *PHASE* (`http://umber.sbs.man.ac.uk/resources/phase/`) For Bayesian divergence time analysis, one can use *BEAST* (`http://evolve.zoo.ox.ac.uk/beast/`).

Software: availability and assistance with analyses

Download the latest version of *MrBayes* here:

`http://mrbayes.csit.fsu.edu/`

A forum for *MrBayes* users is also available:

`http://www.rannala.org/phpBB2/`

An excellent forum for keeping up with the latest developments in phylogenetics:

`http://www.yphy.org/phycom/`

Tracer: a very useful program for visualizing and evaluating the quality of Bayesian MCMC analyses.

`http://evolve.zoo.ox.ac.uk/software.html?id=tracer`

Citations

Rambaut A, Drummond A (2003). Tracer: A program for analysing results from Bayesian MCMC programs such as BEAST & MrBayes, Version 1.3. Distributed by the authors

Ronquist F, Huelsenbeck JP (2003) MRBAYES 3: Bayesian phylogenetic inference under mixed models. *Bioinformatics* **19**, 1572-1574. *(Additional references are listed at the end of this lab handout.)*

Overview of today's lab:

1) Choose partitions and models for your analysis.
2) Add a command block to you NEXUS file to describe your data, model, and MCMC specifications.
3) Run *MrBayes*.
4) Interpret the output.

I. Keep the documentation handy

For any analysis with any software, I recommend first locating and opening the documentation. For help with *MrBayes* you have three options:

1) Download the manual as a PDF file here: `http://mrbayes.csit.fsu.edu/manual.php`
2) Read the manual online: `http://mrbayes.csit.fsu.edu/Help/help.html`
3) Check the Wiki pages: `http://mrbayes.csit.fsu.edu/wiki/index.php/Main_Page`
4) Simply type "help" at the *MrBayes* command prompt.

II. Choosing a model of evolution

Make a new folder for today's lab. Locate your best, most complete NEXUS file containing all your samples, including outgroups, and all your genes, and place a copy of this in your new folder. To help you keep your data and output files organized, I recommend that for each individual analysis you make a new and appropriately labeled folder. Independent *MrBayes* analyses, e.g., assuming different evolutionary models, should be placed in different folders, clearly labeled.

MrBayes requires you to make three decisions regarding your model of DNA sequence evolution.

1) How many partitions will you assume?
2) What model of evolution will be assigned to each partition?
3) Parameters in common among partitions will be linked or independent?

The most basic analysis consists of a single partition containing all the data, *i.e.*, the data are *not* partitioned. The appropriate model is either the same one you selected last week using *Modeltest*, or a more complex model. For students with both nuclear and mitochondrial genes, the two genes should definitely be partitioned, with a different model assigned to each gene (partition). Students with multiple mitochondrial genes may use any of the following approaches:

A) assume one model for all the data, as in *PAUP** (*i.e.*, non-partitioned data).
B) combine protein-coding genes, but partition data by codon position (*i.e.*, 3 partitions).
C) partition protein-coding vs. ribosomal genes (*i.e.*, 2 partitions).
D) partition by codon position and by ribosomal genes (*i.e.*, 4 partitions).
E) partition by codon position and within ribosomal genes, parition loops vs. stems, and assign a "doublet" model to stem positions (*i.e.*, 5 partitions).
F) etc., etc.! You see, there is almost no end to the complexity of your model.

For our first analysis, we will simply assign one model to all the data, *i.e.,* we will simply neglect to partition the data. This model will be the same model, or a slightly more complex model, as the one you chose last week using *Modeltest* and which you assumed for your molecular clock analysis. *MrBayes* recognized the same two parameters to describe heterogeneity among sites in the rates of evolution, I and Γ. However, *MrBayes* allows only three models for the relative rates of the 6 possible types of substitution: JC (1 rate), HKY/F84 (2 rates), GTR (6 rates). If last week you selected a model with either 3, 4, 5 or 6 rates, you should assume the 6-rate model. In Bayesian Phylogenetics, it would be better to assume too many rather than too few models (Nylander et al. 2004; Huelsenbeck & Rannala 2004).

III. Creating your MrBayes block

At the bottom of your new copy of your most complete NEXUS file, add the commands:

```
begin MrBayes ;

END ;
```

All of the following commands will go *inside* this MrBayes block. Note, that *MrBayes* will **not** read any commands you wrote in your `PAUP` blocks. Therefore, any `charset` commands or `taxset` commands will have to be written again in your `MrBayes` block if you want *MrBayes* to be able to use that information, *e.g.*, as a partition or in an "exclude" (characters) or "delete" (taxa) command. Decide now if you need either of these commands, and add them to your `MrBayes` block using the same notation as in your `PAUP` block.

Some basic commands for running an analysis are as follows:

```
begin MRBAYES ;
outgroup 53 ;
set autoclose=yes nowarnings=no ;
lset nst=6 rates=invgamma ngammacat=4 ;
prset shapepr=uniform(0.0001,20.0) ;
prset brlenspr=unconstrained:exponential(10.0) ;
mcmc ngen=20000 nruns=2 printfreq=100 samplefreq=100 nchains=4 Temp=0.10 savebrlens=yes;

END;
```

`begin MRBAYES ;` (Means the following commands will be read by *MrBayes*)

`outgroup 53 ;` (Set the outgroup. Only one sample can be assigned as the outgroup.)

`set autoclose=yes ;` (Less prompting from the program.)

`set nowarnings=no` (Files that already exist will **not** be overwritten. Instead the new output will be appended to the end of the same file. Actually, I prefer `nowarnings=yes` because I keep different runs in different folders anyway, but for first-time uses, the `no` option may be safer.)

`lset nst=6 rates=invgamma ngammacat=4 ;` (nst = number of rate parameters. rates = rate heterogeneity, *e.g.*, `gamma` (Γ), `propinv` (I), `invgamma` (I+Γ). Type "`help lset`" for more information.

`prset brlenspr=unconstrained:exponential(10.0);` (prset = set prior distribution on parameters. If you do not expect super long branches, you can decrease the number to `1.0`, so that the MCMC does not waste time proposing branches that are much too long.)

`prset shapepr=uniform(0.0001,10.0);` (Set prior on Γ-shape parameter. The default is a uniform prior from 0-200, which seems way too wide to me! There seems to be a bug in my version, such that the program will not accept 0.00 as the lower bound of the uniform prior. Strange. You might check your preliminary results and decrease the maximum value still further (perhaps `1.0`).

`mcmc ngen=20000 nruns=2 printfreq=100 samplefreq=100 nchains=4 Temp=0.10 savebrlens=yes;` (`ngens` = number of generations of the MCMC. `nruns` = number of parallel, independent chains to run simultaneously. `printfreq` = how often in MCMC generations the computer screen will be updated. `samplefreq` = how often a sample tree will be saved to the growing collection of trees. `nchains` = total number of parallel Metropolis-coupled MCM chains (MCMCMC). Additional chains experience progressively greater heating. `Temp` = level of heating among chains. Default is 0.20, but lower numbers such as 0.10 may increase mixing by increasing the probability (frequency) of swaps among chains. `savbrlens` = save branch lengths to output files, yes or no? These setting are far too low, but they will serve as a much-too-fast test run.

`END;`

IV. Off and running!

I think in Windows, like Mac OS9, you need to have the data file and the program in the same folder. In UNIX, have your data file in your working directory, rather than giving *MrBayes* long path names leading to your data file. To launch the program, in Windows you will need to use the símolo del sistema again. On a Macintosh computer, you can look for an OS9 version. I much prefer to run *MrBayes* in the UNIX Terminal on my Mac, and I make a hard link of the *MrBayes* executable "mb" to my /bin directory so that I can call it from any other directory.

To launch *MrBayes* type "`mb -i`". The "`-i`" means "interactive mode." Before running your analysis, check for further options you may want to implement by typing at the mb prompt the following commands:
`help lset` = show information on options and settings for the various models of evolution available.
`help prset` = show information on options and settings for prior distributions.
`showmodel` = report the current settings for the model of evolution.
`help mcmc` = show information on options and setting for the MCMC chain.

To start the analysis, type "`execute`" and the name of your data file with the `MrBayes` block of commands, and press return. Your analysis should now be off an running. With the very few generations we are running, your analysis may take 5-30 minutes. Note, the last number in each line of output to the screen is the estimated time until the run ends. The first number on each line is the current generation. Each set of four negative numbers shows the current log-likelihood of each of the four chains in the two independent, parallel analyses. The number in square brackets is the current "cold" chain. Remember that each generation the MCMC makes one proposed change to one parameter. But which parameter and how often? *MrBayes* reports this information for you at the beginning of the run:

```
The MCMC sampler will use the following moves:
With prob.  Chain will change
   4.17 %   param. 1 (revmat) with Dirichlet proposal
   4.17 %   param. 2 (state frequencies) with Dirichlet proposal
   4.17 %   param. 3 (gamma shape) with multiplier
   4.17 %   param. 4 (prop. invar. sites) with sliding window
  62.50 %   param. 5 (topology and branch lengths) with extending TBR
  20.83 %   param. 5 (topology and branch lengths) with LOCAL
```

Your proposal scheme may vary depending on your model.

V. When to stop the analysis?

We want to be sure that the MCMC was sampling from the actual posterior distribution. One way to check this is to run two or more analyses and confirm that they have arrived on the same solution. You have just run two parallel independent MCMC analyses (each with its own heating, as well). Every 1,000 generations *MrBayes* also reports how different are the frequencies of the various clades observed in each of the two independent analyses, *e.g.*, Average standard deviation of split frequencies: 0.073396 Ideally, this number should end up well below 0.01, though we just did a short practice run. When you run your long analysis, be sure this values approaches zero.

After our set number of generations have been run, *MrBayes* will ask us if we would like to continue. If the laboratory session is already half over at this point, you probably should stop the run. On the other hand, if you have time, you can let the analysis run longer.

VI. Summarizing the data.

The above analysis produced one ".t" (tree) and one ".p" (parameter) file for each of the two independent chains. These files contain the data on the trees and parmeters sampled during the MCMC run. We need to summarize the contents of these parameters. We could put the following command (see below) directly into our NEXUS file in the `MrBayes` block, BUT we first need to be sure that our "burn-in" period is adequate. In other words, we do not want to include in our summary those generations of the MCMC before which the chain had reached stationarity. One measure of stationarity is whether the –log-likelihood values have leveled off. To visualize the burn-in of –LnL or model parameters, use the software, *Tracer* version 1.3 (URL listed above), or use *Excel*. With either program open your .p files. In *Excel*, for example, select the second column, labeled LnL. Select the Chart Wizard and make an XY (Scatter) plot. Select the Y-axis and change the minimum and maximum values such that you can zoom in on the curve. Note, the numbers on the X-axis represent the **tree** numbers not the generation numbers. We sampled 201 trees. From my graph, it looks like stationary was probably reached well before tree 100 in both graphs, so I can choose 101

as our burnin value. You have to judge from your own graph. If your results match my number then at the mb command prompt type the *MrBayes* commands:

```
sumt filename=YOUR_filename_here.nex Nruns=2 burnin=101 contype=Halfcompat
```

and press return. `sumt` is the command to summarize the sample of 100s to 100,000s trees you saved in your MCMC run. `filename` = the 'base' of the name of the input file. `Nruns` = the number of parallel, independent analyses you did simultaneously. `burnin` = the number of saved **trees** (not generations) to ignore before you start summarizing the data. `contype` = the type of consensus tree you will want to make. `Halfcompat` is the default and produces as 50% majority-rule consensus tree, but you may want `Allcompat`. *MrBayes* knows to look in `Nruns=2` different files that end in ".t" to estimate the posterior probability distribution of phylogenetic trees.

Next type:

```
sump filename=YOUR_filename_here.nex Nruns=2 burnin=101 Printtofile=YES Outputname=
      YOUR_filename_here.nex.p.stat ;
```

and press return. `Printtofile=YES` means you want the output to go both to the screen **and** to a file, whose name is specified by `Outputname=` , in this case as "`Ag_58_GTRG.nex.p.stat`". This `.p.stat` file will contain only the bottom portion of the information outputted to the screen (*i.e.*, harmonic mean of –LnL values and 95% C.I. of parameter values. You should also select & copy more of the 'standard out' to the screen, starting from the bottom of the screen and including up to the last reported "`Average standard deviation of split frequencies`". Save this data into a new text file, perhaps ending the file name with "`.STDOUT`". *MrBayes* knows to look in `Nruns=2` different files that end in ".p" to estimate the posterior probability distribution of parameter values.

Your Bayesian consensus tree will be in a file ending with ".con". This file will contain two versions of your tree, one with support values and one without. To view the tree, you will have to use *TreeView*, *TreeEdit*, or *PAUP**. To view your tree in *PAUP** you will need to use the MacOS9 version, execute your NEXUS data file in *PAUP**, then use a "Get Trees from File" command. Select "Options..." and "Store branch lengths (if present)". When *PAUP** asks "... Do you want to 'deroot' the tree(s)?", select Yes. Under the Analysis menu, change the criterion to Likelihood, or better yet, in your `PAUP` block add the command "`set criterion = likelihood ;`" so you will never forget to do this step. Now re-root your trees: Trees > Root Tees and select Rooting Options..., then select Make ingroup monophyletic, and select the outgroup topology of your choice. Select Define Outgroup... and chose the appropriate samples, or mid-point rooting if you have no outgroup. When viewing the tree via the "Print Trees..." menu, be sure to click on "Use user supplied branch lengths" (if possible). Save an imagine file (*e.g.*, .pct) of your tree using *TreeView*, *TreeEdit*, or *PAUP**.

VI. Evaluating your MCMC run.

Before executing the longer run, we can compare our prior distributions with our posterior distributions, although after a run of only 20,000 generations, we have a very poor and very unreliable estimate. We can still proceed, with steps **VI.** and **VII.** here, execute a longer analysis, then repeat steps **VI.** and **VII.** as necessary.

Open your output file from the sump command (named, *e.g.*, `YOUR_filename_here.nex.p.stat`). Look at the table of parameter values, showing means, variances, 95% credible interval, etc. Verify that the extremes of the posterior distribution of values for the alpha (α) parameter are still well within the prior uniform distribution you set in the `prset` command in the your `MrBayes` block of your NEXUS file. Note, in this table TL = Tree Length, the sum of all branch lengths in the tree.

You can also check how well your MCMC run was mixing by looking at the frequency at which proposed state changes (remember, one proposal is made each generation) were accepted. Look at your output to the screen for the following information:

```
Acceptance rates for the moves in the "cold" chain of run 2:
    With prob.  Chain accepted changes to
      36.92 %   param. 1 (revmat) with Dirichlet proposal
      12.44 %   param. 2 (state frequencies) with Dirichlet proposal
       7.94 %   param. 3 (gamma shape) with multiplier
      54.59 %   param. 4 (prop. invar. sites) with sliding window
      27.69 %   param. 5 (topology and branch lengths) with extending TBR
      44.23 %   param. 5 (topology and branch lengths) with LOCAL
```

Ideally, these values should fall within 10% and 70%. In the above example, the parameter alpha (α), aka, gamma distribution shape parameter, is not mixing as well. The prior distribution may be too wide or too narrow. (We will assume that the proposal mechanism is fine, and not try to fiddle with that.)

We can also check how well the MCMC is mixing among cold and heated chains. If the heated chains are not occasionally contributing parameter states to the cold chain, then we are wasting CPU time. Often the default heating parameter `T=0.20` is too hot and the more heated chains do not contribute to the exploration and characterization of the posterior probability distribution. At the conclusion of your mcmc command (before sumt or sump commands), you will see a table like this:

```
Chain swap information for run 1:
           1     2     3     4
      -------------------------
   1 |        0.38  0.20  0.00
   2 |  3415        0.50  0.01
   3 |  3253  3281        0.09
   4 |  3374  3404  3273
```

This table shows the number of proposed state swaps between chains (below the diagonal) and the frequency at which the proposal was accepted (above the diagonal). When two chains swap states it means that the more heated chain was at a point in state space (the combined topology, branch lengths, and evolutionary model parameters) with a better log-likelihood score than the less-heated (or cold) chain, and all states are then swapped together. The above frequencies are adequate. If your frequencies are lower, you should lower the Temp parameter for the next run. Alternatively, if you see that the forth chain is not contributing to the MCMCMC analysis, and you find that your analysis is taking much too much time, you might run your analysis with just three chains. If your swap frequencies are much higher than those shown above, you might try raising the Temp parameter, perhaps restoring the default value of `T=0.20`.

VII. Executing a longer MCMC run.

Now you need to run a proper analysis involving millions of generations, with a much lower sample frequency. You might try the following settings:

```
mcmc ngen=2000000 nruns=2 printfreq=1000 samplefreq=1000 Diagnfreq=10000 nchains=4
            Temp=0.10 savebrlens=yes ;
```

With these settings, you would sample (`2000000/1000) x 2 chains` or 4,001 trees. If your burnin period of, *e.g.*, 501 trees, your final estimate of the posterior distribution would be based on 3,500 trees. If your analysis is taking too much time, you might use just 3 chains. Ideally, you would not want to run fewer generations, but for purposes of this lab you might have to, in which case you might also sample less frequently, *e.g.,* every 500 generations. Keep in mind, however, that if only 4.17% of the generations involve a change in the alpha parameter, and only ~10% of these are accepted, then we have to wait on average ~250 generations between successful updates to that parameter. After 500 generations the parameter value would have changed only twice, and some autocorrelation among samples would result. If you find that the above analysis is running quickly, you might stop your analysis and add 1 or 2 million more generations, and then restart the MCMC run. Your goal is to obtain as many samples as possible from the posterior distribution while at the same time insuring that the samples are independent. Because this is a "Monte Carlo" method, we need large sample sizes to estimate well the posterior probabilities.

When this run ends, return to steps **V.** and **VI.** above to remind yourself how to summarize the data and observe your Bayesian consensus tree. Be sure to save your .con file and also save your Bayesian consensus tree as an image file (e.g., pct file) for possible inclusion in your final manuscript.

X. Partitioning your data

If you want to partition your data by gene or by codon position, here is an example. Below, we first define some character sets, *e.g.*, genes or codon positions. We then define one or more partition schemes. Note, only partition can be active at a time. We "activate" a partition using the `set partition` command (see below). In example below,

we have defined one partitions by gene (cyt *b* vs. COI) and one by position (first vs. second vs. third), but to partition the data 6 ways, we would have to define a new, third partition. Next we apply different evolutionary models to each partition using the `lset applyto=` command. In the example below, we are applying a GTR+I+Γ model to the first position, the second position will assume a HKY+I model, the third position will assume a GTR+Γ model. Note, if you want the nucleotide frequencies to be fixed at equal frequencies, you need to change the prior distribution by implementing the following command to that partition: `prset statefreqpr=fixed(equal)`. Finally, we use the `unlink` command to declare that when the same parameter is observed in more than one of the subpartitions, we want that parameter to be estimated independently in each. All of these commands are written inside the `MrBayes` block before the `mcmc` command.

```
charset COI = 1-639 ;
charset cytb = 640-1356 ;
    charset first_pos  = 1-1354\3 ;
    charset second_pos = 2-1355\3 ;
    charset third_pos  = 3-1356\3 ;
partition genes = 2:COI,cytb ;
partition by_codon = 3:first_pos,second_pos,third_pos ;
set partition=by_codon ;
lset applyto=(1) nst=6 rates=invgamma Ngammacat=5 ;
lset applyto=(2) nst=2 rates=propinv ;
lset applyto=(3) nst=6 rates=gamma Ngammacat=5 ;
unlink Statefreq=(all) ;
unlink Shape=(1,3) ;
unlink Pinvar=(1,2) ;
unlink Revmat=(1,3) ;
```

Lab Report:

To your final report you will need to add the results of *MrBayes*. You should either include your ML tree or your Bayesian consensus tree, but in either case you should report the marginal posterior probabilities for each node (clade), or at least note which nodes received at least 0.95 probability of being correct. Use a graphics program to make a nice phylogenetic tree figure. You may need to add in the posterior probabilities by hand. Update your Methods sections and your Results section. Expand your Literature Cited section as necessary.

Please answer the following questions, and hand in your answers along with your beautiful Bayesian consensus phylogram figure (graphic) with bipartition probabilities (support values), scale bar, and OTUs with informative labels. Also indicate such information as taxonomy or localities in your tree. These labels may go to the right of your tree, if you wish.

(1) What were the search conditions of your longer Bayesian MCMC run? List the number of generations, the number of chains per run, the number of independent parallel runs, and the sampling frequency. How many trees did you discard as your burn-in period and how did you decide on this number? How many trees were used to calculate your posterior probabilities and consensus tree? Which parameter had the lowest update acceptance rate during your MCMC and what was its update frequency?

(2) How does your Bayesian consensus tree compare to your ML tree from last week?

(3) Describe two potential problems or "pitfalls" with Bayesian MCMC phylogenetic analysis. Describe how you attempted to avoid (or at least evaluate the severity of) these potential problems in your own analysis. Your answer should be about a paragraph in length.

Some useful references on Bayesian phylogenetics:

Castoe TA, Doan TM, Parkinson CL (2004) Data partitions and complex models in Bayesian analaysis: the phylogeny of gymnphthalmid lizards. *Systematic Biology* **53**, 1-22. *(An empirical exploration and comparison of alternative strategies to data partitioning and their effect on inferred levels of clade support.)*

Erixon P, Bodil B, Britton T, Oxelman B (2003) Reliability of Bayesian posterior probabilities and bootstrap frequencies in phylogenetics. *Systematic Biology* **52**, 665-673.

Felsenstein J (2004) *Inferring Phylogenies*. Sinauer Associates Inc., Sunderland, Massachusetts, USA. *(Chapter 18 provides an excellent introduction to Bayesian phylogenetics and its potential pitfalls.)*

Holder M, Lewis PO (2003) Phylogeny estimation: traditional and Bayesian approaches. *Nature Reviews Genetics* **4**, 275-284. *(An introduction to the differences between Bayesian and likelihood [frequentist] methods.)*

Mossel E, Vigoda E (2005) Phylogenetic MCMC algorithms are misleading on mixtures of trees. *Science* **309**, 2207-2209. *(Combining genes with truly different histories can be invalidate Bayesian MCMC analysis.)*

Huelsenbeck JP, Rannala B (2004) Frequentist properties of Bayesian posterior probabilities of phylogenetic trees under simple and complex substitution models. *Systematic Biology* **53**, 904-913. *(Over-parameterization is better than under-parameterization in Bayesian phylogenetics.)*

Huelsenbeck JP, Ronquist F, Nielsen R, Bollback JP (2001) Bayesian inference of phylogeny and its impact on evolutionary biology. *Science* **294**, 2310-2314.

Lewis PO (2001) A likelihood approach to estimating phylogeny from discrete morphological character data. *Systematic Biology* **50**, 913-925. *(Model-based phylogenetic analysis of morphological data.)*

Lewis PO (2001) Phylogenetic systematics turns over a new leaf. *Trends in Ecology and Evolution* **16**(1), 30-37. *(Yet another review article on Bayesian phylogeny.)*

Nylander JAA, Ronquist F, Huelsenbeck JP, Nieves-Aldrey JL (2004) Bayesian phylogenetic analysis of combined data. *Systematic Biology* **53**, 47-67. *(Partitioning your data to increase model complexity – advantages and advice on using Bayes Factors to evaluate non-nested alternative models.)*

Pagel M, Meade A (2004) A phylogenetic mixture model for detecting pattern-heterogeneity in gene sequence or character-state data. *Systematic Biology* **53** (4), 571–581. *(***BayesPhylogenies** *software: The uses choices the number of partitions, and the software figures how to spread these partitions among your data!)*

Wilcox TP, Zwickl DJ, Heath TA, Hillis DM (2002) Phylogenetic relationships of the dwarf boas and a comparison of Bayesian and bootstrap measures of phylogenetic support. *Molecular Phylogenetics and Evolution* **25**, 361–371. *(Comparing levels of support obtained from Bayes and bootstrap methods.)*

Yang Z, Rannala B (2005) Branch-length prior influences bayesian posterior probability of phylogeny. *Systematic Biology* **54**(3), 455–470. *(Like the title says...)*

Additional important citations are obtained by typing the command "citations" at the MrBayes command prompt.