# Markov chain Monte Carlo II

Fredrik Ronquist

October 12, 2005

## 1 Introduction

In this lecture, we will be covering some of the most common types of moves used in Bayesian phylogenetic inference. We will also talk about the Green extension to MCMC methods, also known as reversible jump MCMC. Finally, we will briefly cover convergence monitoring.

## 2 The MCMC algorithm in a nutshell

The most important step in the MCMC algorithm is the calculation of the acceptance probability $r$. As we have seen, it is the product of three ratios: the prior ratio, the likelihood ratio, and the Hastings ratio. If we are considering a move from parameter values $\theta$ to $\theta'$ according to a proposal density $q(\theta'|\theta)$, we should accept the move with the probability

$$r = \min\left\{1, \frac{p(\theta')}{p(\theta)} \times \frac{p(X|\theta')}{p(X|\theta)} \times \frac{q(\theta|\theta')}{q(\theta')|\theta)}\right\}$$

Usually, this product is calculated easily, enabling one to run the Markov chain for many generations.

# 3 Choice of proposal mechanisms

Convergence to the stationary distribution is ensured for a wide variety of proposal distributions. The only requirement is basically that the proposal mechanism allows the chain to go from any point in parameter space to any other point in a finite number of steps (with probability larger than 0).

In practice, however, some proposal mechanisms converge much faster than others. Typically, a proposal mechanism will have a tuning parameter, which determines how bold the proposals are. This tuning parameter is then set such that the proposal is not too modest or too bold. If the proposals are too modest, they will always be accepted but it will take a long time to cover the region of high probability mass in the posterior distribution. If the proposals are too bold, most of them will be rejected and it will again take a long time to cover the region of interest.

For multi-parameter models, there are several ways of structuring the proposals. A common way is to update one parameter or a block of parameters at a time. This usually leads to fast calculation of the prior and likelihood ratios. Single-parameter updates work well for parameters that are uncorrelated. If there is strong correlation between two parameters, it is typically advantageous to propose new values for both of the correlated variables in the same Metropolis step, particularly if the proposal can be made in the direction of the correlation. For instance, if parameter $a$ is positively correlated with parameter $b$, we may want to propose a decrease of $a$ together with a decrease of $b$, and an increase of $a$ together with an increase of $b$. This requires, however, that we know something about the shape of the posterior distribution. In some problems it is possible to avoid correlation between parameters by choosing an alternative parametrization, and this is often recommended.

If we use a collection of proposal mechanisms for a multi-parameter model, then you can either cycle through the proposals systematically or randomly choose one proposal in each step of the Markov chain. Both approaches are used in practice.

We will illustrate the design of MCMC proposal mechanisms with a few examples from Bayesian phylogenetic inference.

# 4 Sliding window proposal

Assume the tree is given and that we want to sample the posterior probability distribution over branch lengths. The prior on individual branch lengths is an exponential distribution with parameter $\lambda$, that is $v \sim \text{Exp}(\lambda)$. This means that $v$ has the density

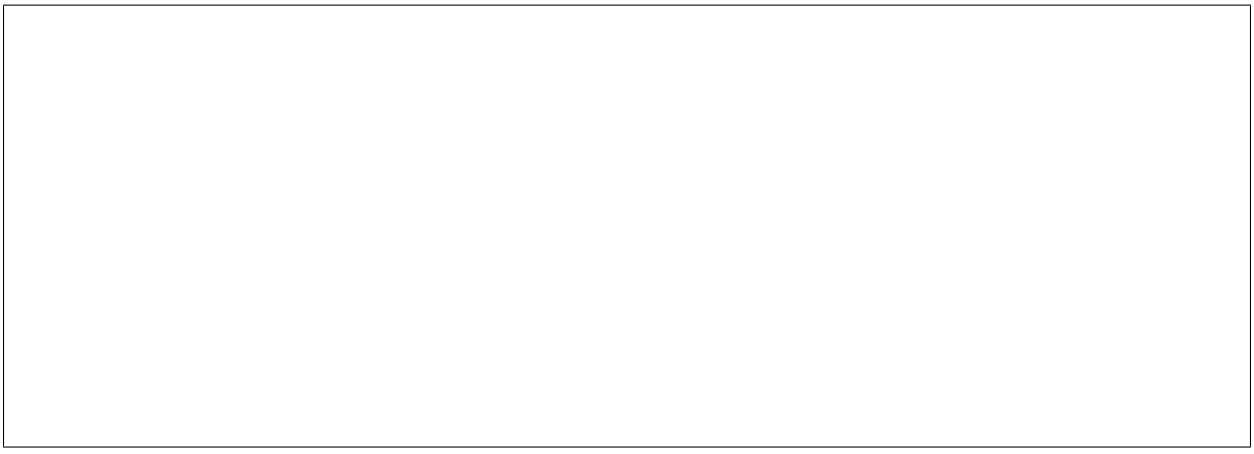
$$f(v) = \lambda e^{-\lambda v}$$

As we have stated earlier, we can start with arbitrarily chosen branch lengths; the chain will converge onto the posterior probability distribution of branch lengths regardless of the starting values. A simple approach we might use for changing these initial branch length values is to have our proposal mechanism randomly pick one branch in the tree and then change that branch length with a *sliding window proposal*. Such a proposal centers a window on the current value $v$. The window has the width $\delta$, which is a tuning parameter. A random number $u$ is drawn from a uniform distribution on the interval $(0, 1)$ (that is, $f(u) = 1$), and the new value for the branch length, $v'$ is obtained as

$$v' = v + (u - 0.5)\delta$$

If we think about the proposal graphically, we are simply proposing new values uniformly within the sliding window centered on the current value (Fig. 1).

Figure 1: Sliding window proposal. New values are chosen uniformly from a window centered on the current value.



How do we calculate $r$? The prior ratio is obtained from the density of the exponential distribution:

$$\frac{p(v')}{p(v)} = \frac{\lambda e^{-\lambda v'}}{\lambda e^{-\lambda v}} = e^{\lambda(v-v')}$$

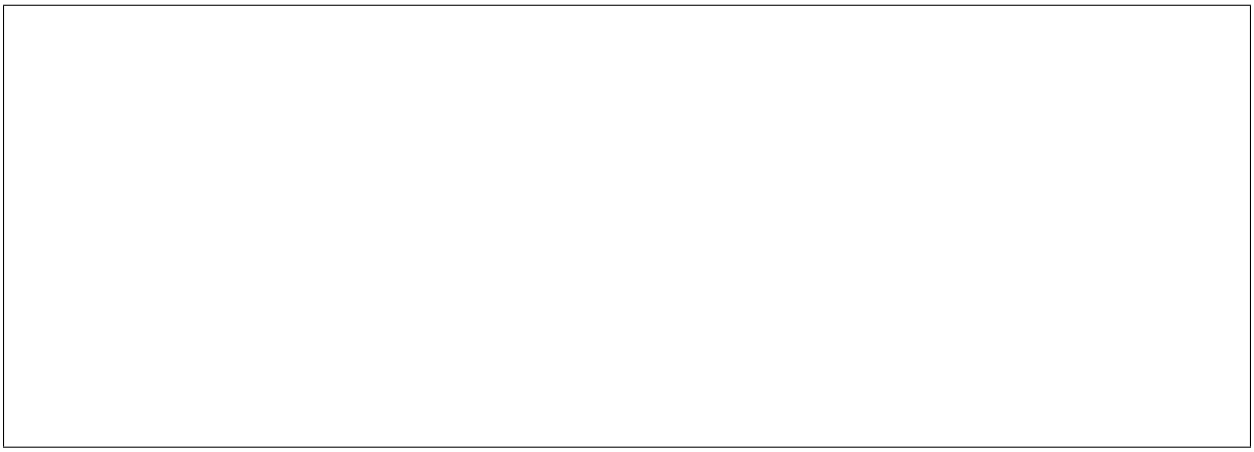
The likelihood ratio is obtained by calculating the likelihood of the tree with the old branch length and the likelihood of the tree with the new branch length. If we calculate the conditional likelihoods of all states of both subtrees attached to the branch, then this likelihood ratio can be obtained simply by computing the likelihood over the branch using two different transition probability matrices; in practice, we would store the likelihood of the current tree and only calculate the likelihood given the new proposed branch length.

Finally, how is the Hasting's ratio calculated? For the sliding window proposal, intuitive reasoning may be sufficient. Say that moving from $v$ to $v'$ requires picking the random number $u$. Then it is easy to see that the reverse move, going from $v'$ to $v$, requires picking the random number $1 - u$ (these values will be symmetric around 0.5). Since we are picking the random numbers from a uniform distribution, these probabilities will be the same and the Hastings ratio will be 1.

One final point deserves some discussion. What do we do when we reach the boundaries of parameter space? Theoretically, we can define an exponential prior on branch lengths on the interval $(0, \infty)$, but in practice we typically restrict this interval to $(a, b)$ where $a$ and $b$ are machine constants determining the interval of values that can be handled with reasonable precision and speed on the machine we are using. When the current value is close to one of these machine constants, then straight-forward application of the sliding window proposal will result in some values being proposed outside of the valid interval. There are two solutions to this problem. The most elegant solution is to use reflection at the boundary. Say that we would have proposed a new value $v' > b$ using the standard sliding window. Then we reflect that value back in to produce the new proposed value $v'' = b - (v' - b)$. The Hastings ratio for the reflected values is 1; there will be two ways of going between two values near a boundary but those two possibilities are always the same for both directions of the move (Fig. 2).

An alternative solution that works equally well is to abort proposals that will take you outside of the permissible interval. You need to always accept the old state when you abort the proposal; this compensates for the lack of proposals coming back from the values outside of parameter space to the values close to but inside the boundary.

Figure 2: Reflection with the sliding window proposal. New values that are outside of a boundary are reflected back in.



## 5 Multiplier proposal

A disadvantage with the sliding window proposal is that it moves very slowly between large branch length values. Thus, if we started with very long branches and the bulk of the posterior probability distribution was on small branches, it might take a long time before the chain converges. Say that the starting value is 100 and that most of the posterior density is centered around 0.01. If we use a sliding window with the width 0.1, then it will take a very long time before the chain finds the region of high posterior density. Not only does this require a large number of small steps; it is also likely that the posterior probability distribution is virtually flat around the value of 100, giving the Markov chain no directional guidance. The problem cannot be solved by simply increasing the window size because then the chain will suffer from slow mixing once it has reached the region of interest because it will make too bold proposals.

A proposal that deals better with wide ranges of parameter values than the sliding window is the *multiplier proposal*. The idea is to multiply the current value with a number drawn from a suitable distribution. A common choice is to draw the multiplier $m$ from the distribution with the density

$$f(m) = \frac{1}{\lambda m} \quad : \quad a < m < b$$

Specifically, the values $a$ and $b$ are chosen such that $a = 1/b$. Thus, we will generate multipliers that will maximally multiply or divide the current value with the same boundary value $b$. For this distribution to normalize correctly, we need to set $\lambda = 2\ln(b)$ as we will see below.

To draw randomly from this distribution, we use the observation that the primitive function of the density function, $F(m)$ is distributed as a uniform on the interval $(0, 1)$. Thus, by drawing a random number on the interval $(0, 1)$ and then calculating the inverse of $F(m)$, we can generate random variables with the density function $f(m)$. In our case, the general-form primitive function is

$$F(m) = \frac{\ln(m)}{\lambda} + C$$

If we add the boundary conditions that $F(1/b) = 0$ and $F(b) = 1$ we get

$$F(m) = \frac{\ln(m)}{2\ln(b)} + \frac{1}{2}$$

If we draw a random number $u$ from a uniform on $(0, 1)$ and then set $F(m) = u$, we can easily solve for $m$ (using $\lambda$ instead of $2ln(b)$ for brevity):

$$\begin{aligned} u &= \frac{\ln(m)}{\lambda} + \frac{1}{2} \\ \ln(m) &= \lambda(u - 0.5) \\ m &= e^{\lambda(u-0.5)} \end{aligned}$$

Suppose we use this mechanism to update a branch length $v$ to a new value $v' = mv$. The prior and likelihood ratios can be calculated as before but the Hastings ratio is more difficult. Here, we will cover a method described by Green (1995), which is based on separating the Hastings ratio into two components, one which deals with the generation of the random numbers for the forward and backward move, and one which deals with any stretching or shrinking of the parameter space that occurs when we go from one state to the other.

Assume that we draw one or more random numbers $u$ from the distribution $g$ and form the proposed state $x'$ by some deterministic function of the current state $x$ and the random numbers, say $x' = h(x, u)$. To go back, we would draw the random numbers $u' \sim g'$ to give $x = h'(x', u')$. The Hastings ratio $q(x|x')/q(x'|x)$ can now be calculated as a product of simple densities for the random numbers and the Jacobian describing the stretching or shrinking during the move:

$$\frac{q(x|x')}{q(x'|x)} = \frac{g'(u')}{g(u)} \left| \frac{\partial(x', u')}{\partial(x, u)} \right|$$

Let $u = m$ be the multiplier needed to go from branch length $x = v$ to $x' = v' = mv$. We have that $x' = h(x, u) = ux$. To go back from $x'$ to $x$ we need the multiplier $u' = 1/u$, and we have $x = h'(x', u') = (ux)(1/u) = x$. The density functions $g$ and $g'$ are identical; they are the density function given above for the multiplier. The Hastings ratio can now be calculated using Green's method as:

$$\begin{aligned}
\frac{q(x|x')}{q(x'|x)} &= \frac{g'(u')}{g(u)} \left| \frac{\partial(x', u')}{\partial(x, u)} \right| \\
&= \frac{\frac{1}{\lambda(1/u)}}{\frac{1}{\lambda u}} \begin{vmatrix} \frac{\partial x'}{\partial x} & \frac{\partial x'}{\partial u} \\ \frac{\partial u'}{\partial x} & \frac{\partial u'}{\partial u} \end{vmatrix} \\
&= u^2 \begin{vmatrix} \frac{\partial ux}{\partial x} & \frac{\partial ux}{\partial u} \\ \frac{\partial (1/u)}{\partial x} & \frac{\partial (1/u)}{\partial u} \end{vmatrix} \\
&= u^2 \begin{vmatrix} u & x \\ 0 & -u^{-2} \end{vmatrix} \\
&= u^2 \left| (-u^{-1} - 0) \right| \\
&= u
\end{aligned}$$

Thus, the Hasting's ratio simplifies to $u$. It is relatively easy to understand why the Hasting's ratio must be $u$ by simply considering how values stretch or shrink when going from a branch length of $x$ to a length of $x'$ (Fig. 3).

## 6 The LOCAL move

To go between different trees in a Bayesian MCMC analysis of phylogeny, it is not sufficient to change branch lengths; we also need to change topology. We will cover one topology proposal here, namely the LOCAL move, one of the first topology proposals used for Bayesian phylogenetic analysis (Larget and Simon, 1999). The proposal changes both branch lengths and (potentially) topology, and can thus be used as the only proposal mechanism for Bayesian MCMC analysis of phylogeny under the Jukes Cantor model.

The LOCAL move first selects three adjacent branches in the tree, the middle of which must be internal (Fig. 4). Call the lengths of these branches $a$, $b$ and $c$. Each of the branch lengths is now hit with the same multiplier $m$, selected as described above. Thus, we get the new branch

Figure 3: Interpretation of the Jacobian for the multiplier proposal. Note that multiplying a value by a factor larger than 1 results in thinning of the values for any given numerical precision.

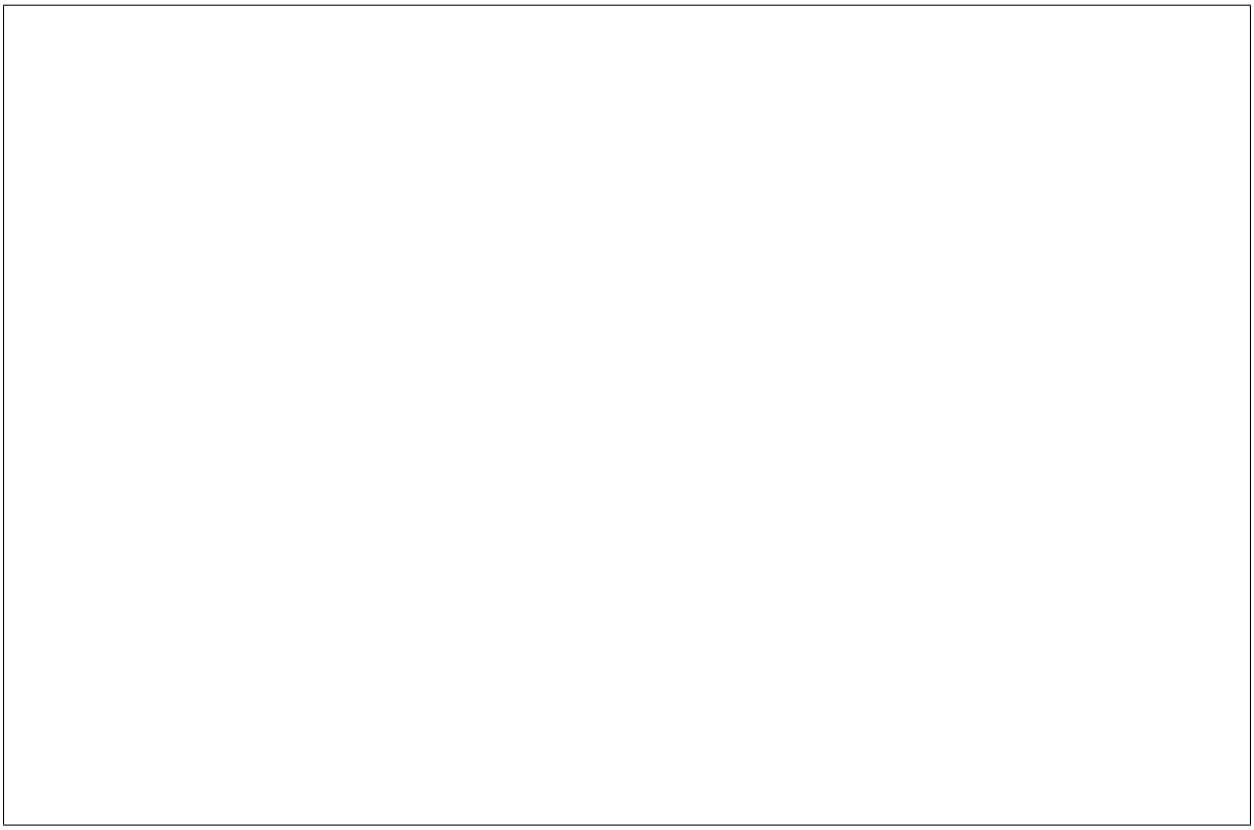
lengths $ma$, $mb$, and $mc$. Finally, one of the two subtrees attached to the three-branch piece is chosen at random. This subtree is removed and then reinserted with uniform probability over the three-branch piece. The removal and reinsertion will cause a change in branch length proportions and, in some cases, also a change in topology (Fig. 4).

Calculating the Hasting's ratio for the LOCAL move is relatively difficult, and the originally cited Hasting's ratio is incorrect. We can use Green's method described above on the entire proposal but it is easier to divide it into separate parts, noting that the Hastings ratio for the entire move must be the product of the Hastings ratios of the independent parts. In the first part of the move, we simply multiply three different branch lengths with the same multiplier. The Hasting's ratio for this part of the move is $m^3$. The second part of the proposal changes branch lengths by adding and subtracting pieces between branch lengths; the Hasting's ratio for this part is 1, so that the total Hasting's ratio for the local is $m^3$.

## 7 Reversible jump MCMC

An interesting extension of MCMC techniques, due to Green (1995), is referred to as reversible jump MCMC. This technique is used to jump between models with different numbers of parameters. The moves for these MCMC chains are set up in a similar way to those for regular MCMC chains.

Figure 4: The LOCAL proposal. Three adjacent branches are chosen, of which the middle one must be internal. The three branches are then multiplied by the same multiplier value. Finally, one of the two subtrees attached to the three-branch segment is chosen randomly and reinserted with a uniform probability on the three-branch segment.



Green's method for calculating Hasting's ratios still holds; the vectors $x$ and $x'$ do not need to have the same number of elements, as long as the vectors $(x, u)$ and $(x', u')$ do. For instance, we can move from an $x$ with one parameter to an $x'$ with two parameters using two random numbers, as long as we can move in the other direction using only a single random number. In this case, both $(x, u)$ and $(x', u')$ would contain three values.

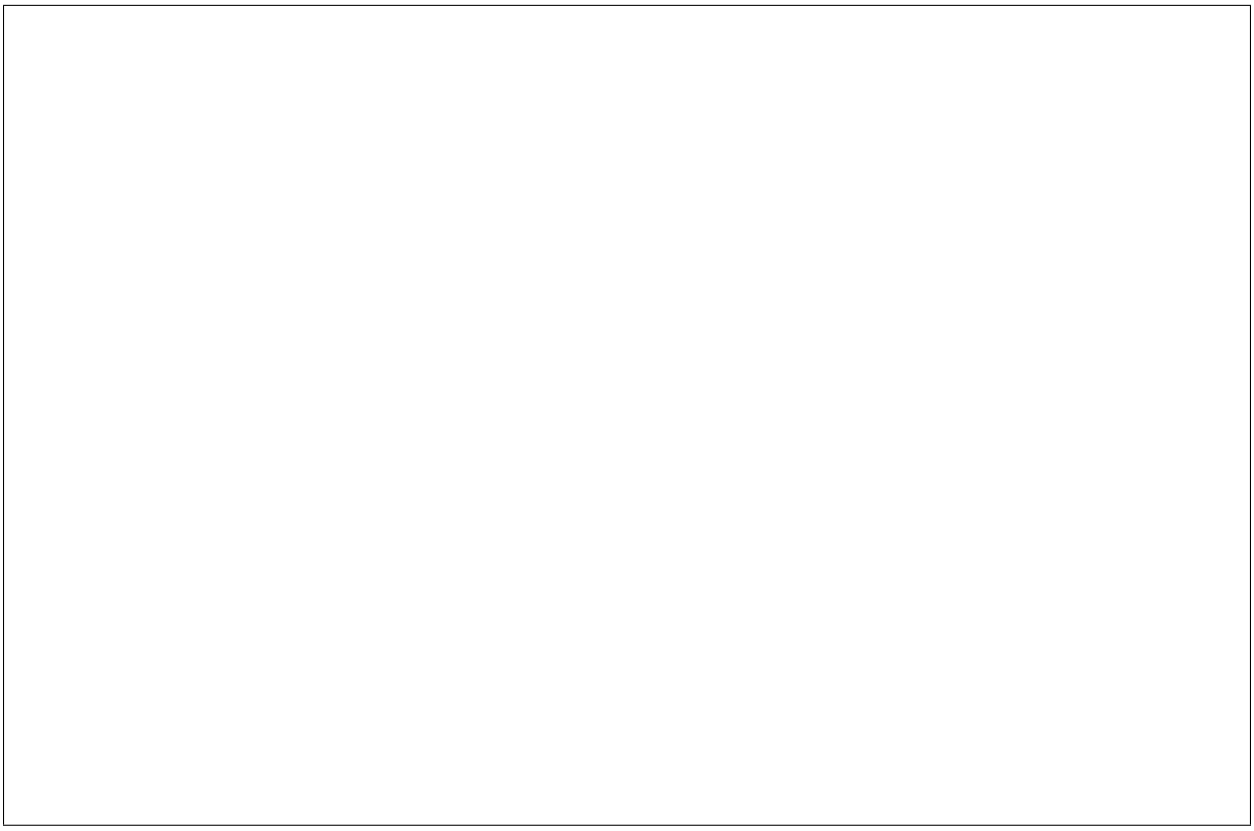
A simple example of a model needing reversible jump methods is the Compound Poisson process. When using MCMC methods to explore this model, we need to add and remove events, moves that change the dimensionality of the system.

# 8 Convergence diagnostics

MCMC methods are guaranteed to converge onto the target distribution if run long enough. The difficulty, however, lies in determining how long is long enough. There is a great deal of interest, therefore, in methods for assessing the convergence of Markov chains.

People generally take three different approaches to convergence assessment. The simplest and fastest method is to look at the plot of likelihood values over generations of the chain, the so called *trace plot*. When the likelihood values reach a stable plateau and stop increasing from generation to generation, then the chain may have converged (Fig. 5). Unfortunately, however, trace plots are notoriously unreliable as indicators of convergence. If the trace plot has not stabilized, then it is unlikely that the chain has converged, but the reverse is not necessarily true. Therefore, trace plots *must* be complemented with other methods of assessing convergence.

Figure 5: Different approaches to convergence diagnostics.



One possible approach to convergence monitoring is to compare samples within one very long run. If the run has converged onto the stationary distribution, then it should be true that subsections of the final part of the run have sampled similar values. Thus, by comparing such subsections

(windows) of the run, we may detect problems with convergence (Fig. 5).

Convergence monitoring by comparison of values between windows of a very long run is better than just examining trace plots but it is still relatively risky. It is easy to construct examples where both trace plots and single-run convergence diagnostics seem to indicate convergence long before any reasonable sample from the posterior has been obtained. An alternative and safer method of convergence checking is to compare different runs started from different starting values (Fig. 5). In the beginning of the run, such chains will sample very different regions of parameter space, but as they converge onto the stationary distribution, the samples will look more and more similar to each other.

For scalar parameters, there are various indicators that summarize such a comparison across multiple runs. Perhaps the most popular indicator is Gelman and Rubin's statistic, the *Potential Scale Reduction Factor*, which is a comparison of the variance between and within runs. For more complex parameters, it is necessary to design appropriate indicators. We have used the average standard deviation of clade support values as a rough guide to convergence of MCMC runs on phylogeny problems; it seems to work well as a convergence indicator.

## 9 Study Questions

1. Describe the sliding window proposal
2. Why can the sliding window proposal be inefficient for sampling from some posterior distributions?
3. Derive the Hasting's ratio for the sliding window proposal
4. Compare and contrast different approaches to convergence monitoring
5. When is reversible-jump MCMC used?