# Bayesian Inference using Markov Chain Monte Carlo in Phylogenetic Studies

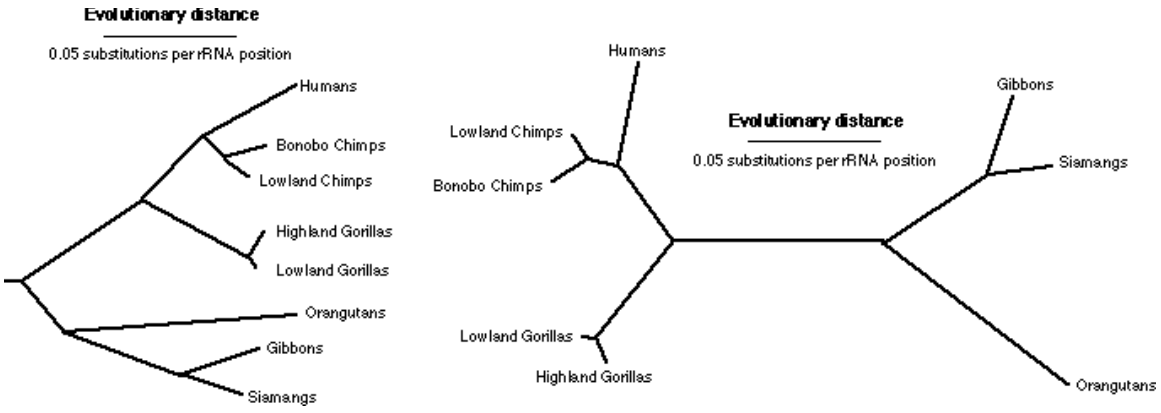
Essay written for the course in Markov Chains 2004
Torbjörn Karfunkel

## 1 What is phylogeny?

Phylogeny is the evolutionary development and history of a species or higher taxonomic grouping of organisms. Much effort is put into reconstructing phylogenies, since understanding evolutionary relationships is needed to draw general conclusions about the mechanisms of evolution.

One must be aware that much of the terminology used in biology is quite vague – concepts such as 'species', 'individual', 'organism', and even 'life' are arbitrary terms whose definitions may vary. Some of the entities studied are not regarded as organisms, e.g. viruses. Still, the generally accepted conception within the scientific community is that all these entities have a common ancestor. Ancestry in this case refers to the passing on of genetic material, DNA and RNA. These substances carry information on how to build and operate an organism, and they are copied and passed on to offspring. However, the copies made are often not perfect. *Mutations* occur, and it is the accumulated weight of these mutations along with natural selection that constitute the driving forces of evolution. In other words, 'life' has arisen on earth once only, and from this origin all other life forms have evolved.

The conception that all life forms have a common origin has led to the use of trees (usually binary) for representing the interrelationships among organisms. At the root is the *most recent ancestor*, and at the tips are the organisms studied. Branch lengths correspond to time, or sometimes to *evolutionary distance*, meaning the mutational changes that have occurred.




Two trees in which the branch lengths represent evolutionary distance. The only difference between these two trees is that the one on the left is rooted, whereas the one on the right is not. In a rooted tree, the root corresponds to the most recent ancestor.

## 2 Defining the problem

The main task for phylogenetic research is to reconstruct evolutionary history. Some clues are provided by fossils, but it is often the case that such evidence is absent, and in these cases conclusions must be drawn using only the features that are observable in present-day organisms. The only way to go about in this situation is to use similarities between the organisms. Similarities are often indications of close relations.

Before the boost of molecular data, phylogenetic analysis was usually performed on visible physical features. Nowadays the starting point of the analysis is mainly sequence information – either protein or DNA/RNA sequences. The high availability of this reliable source of evidence does not, however, imply the end of concerns for phylogenetic research. There are still computational aspects that constitute obstacles. Aligning sequences in order to determine which similarities and differences exist is in itself difficult, but not within the scope of this essay. Instead it is assumed here that the sequences have been aligned in a credible manner.

One of the most apparent problems is the vast number of possible tree topologies. For unrooted trees the number of possible topologies for *s* organisms is

$$\frac{(2s-5)!}{2^{s-3}(s-3)!}$$

whereas the corresponding number for rooted trees is

$$\frac{(2s-3)!}{2^{s-2}(s-2)!}$$

These numbers grow rapidly with increasing values of *s*, thus making extensive analysis intractable (Huelsenbeck and Ronquist, 2001).

| s | Unrooted trees | Rooted trees |
|---|---|---|
| 4 | 3 | 15 |
| 5 | 15 | 105 |
| 6 | 105 | 945 |
| 7 | 945 | 10,395 |
| 10 | 2,027,025 | 34,459,425 |
| 20 | $2.22 \cdot 10^{20}$ | $8.2 \cdot 10^{21}$ |

### 2.1Models of sequence evolution

The statistically more robust methods for reconstructing phylogenies attempt to compute probabilities of different phylogenetic trees. In order to do so, they must incorporate a *model of sequence evolution.* Such a model takes into consideration the facts that not all mutations are equally probable in a sequence.

Many such models exist, with varying levels of abstraction and complexity. In general, the more realistic the model is, the more parameters are included, and as a result the computational burden will increase. Normally the user of a software for phylogenetic analysis only has to specify which model to use, and the parameters involved will be calculated during run-time, in a fashion that maximizes the probabilities sought for, based on the sequence information that constitutes the input of the analysis.

## 3 Some approaches

Several approaches have been used to infer phylogeny from data, and some of them will be briefly outlined here.

### 3.1 Neighbor joining

The *neighbor joining* method takes as its input a *distance matrix*, in which all pairwise distances between sequences are listed. A *greedy algorithm* is then used to pairwise join the two closest neighbors until the result is a binary tree. No model of sequence evolution is needed.

### 3.2 Maximum parsimony

The idea behind *maximum parsimony* is to construct a tree in which the sum of all mutations along all branches is minimized. No model of sequence evolution is needed.

### 3.3 Maximum likelihood

*Maximum likelihood* is a very robust approach. It tries to find the tree for which the likelihood is maximized, among all possible tree topologies and parameters. A model of sequence evolution is needed. Unfortunately the number of possible trees is so vast that extensive search is not feasible even for a modest number of organisms. Therefore the search is usually done using a *hill-climbing* algorithm, always looking for changes to the tree that will increase the likelihood. The search is often started using a tree computed by a simpler method, such as neighbor joining.

### 3.4 Establishing confidence using bootstrapping

All of these methods will give as a result a single phylogenetic tree (the best estimate). Normally you want a measure of confidence too. How confident are we that a certain group of organisms belong to the same branch? Within which boundaries can we be 95% sure that this branch length lies?

In order to establish such confidence limits *bootstrapping* is used. In bootstrapping a number of *pseudoreplicates* are constructed by random resampling, with replacement, from your aligned sequences. The pseudoreplicates are of the same length as the original aligned sequences. The same tree-building method as earlier is then used on these pseudoreplicates, and from this set of trees confidence limits can be computed. Obviously this approach will increase the computational burden by a large factor – if 100 bootstrap samples are used, the computation will take 101 times as long. For the fast methods this is not a big problem, but for maximum likelihood, that is already computationally intense, the extra complexity may well be more than one can afford.

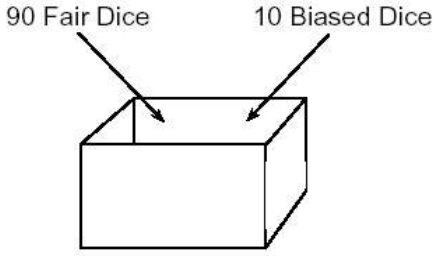
## 4 Bayesian Inference

In recent years *Bayesian inference* has been introduced as a powerful method to reconstruct phylogenies. Bayesian inference is in a way connected to maximum likelihood, but the basic question one asks is completely different.

In maximum likelihood the question is *'What is the probability of seeing the observed data (D) given that a certain model/assumption (T) is true, P(D|T)?*', whereas the question in Bayesian inference is *'What is the probability that this model/assumption (T) is correct, given the data (D) that we have observed, P(T|D)?*' The latter question is answered by *Bayes theorem*:

$$P(T|D) = \frac{P(T)P(D|T)}{P(D)}$$

P(T|D) is the *posterior probability*.
P(T) is the *prior probability*, that is the probability assigned to the theory T *before* the data was collected.
P(D|T) is the *likelihood*.
P(D) is the probability of observing the data D disregarding which theory is correct.
An example:




| Observation | Fair | Biased |
|---|---|---|
| ⚀ | 1/6 | 1/21 |
| ⚁ | 1/6 | 2/21 |
| ⚂ | 1/6 | 3/21 |
| ⚃ | 1/6 | 4/21 |
| ⚄ | 1/6 | 5/21 |
| ⚅ | 1/6 | 6/21 |

90 fair and 10 biased dice are put into a box, and a die is then drawn at random from the box. What is the probability that the die is biased?
P(Biased) = 10/(90+10) = 0.1 This is our prior probability!
Now assume that we roll the die twice, once rolling a ⚃ and once rolling a ⚅. What do we think now about the die?
Using the maximum likelihood approach we get that
P(⚃ ⚅ | Fair) = 1/6 × 1/6 = 1/36

P(⚃ ⚅ | Biased) = 4/21 × 6/21 = 24/441

Since P(⚂ ⚅ | Biased) > P(⚃ ⚅ | Fair) we draw the conclusion that the die is biased.

Using the Bayesian approach we get

$$P(\text{Biased} \mid \boxdot\boxdot) = \frac{P(\text{Biased}) \times P(\boxdot\boxdot \mid \text{Biased})}{P(\boxdot\boxdot)}$$

This is our prior, 0.1

This is the likelihood, 24/441

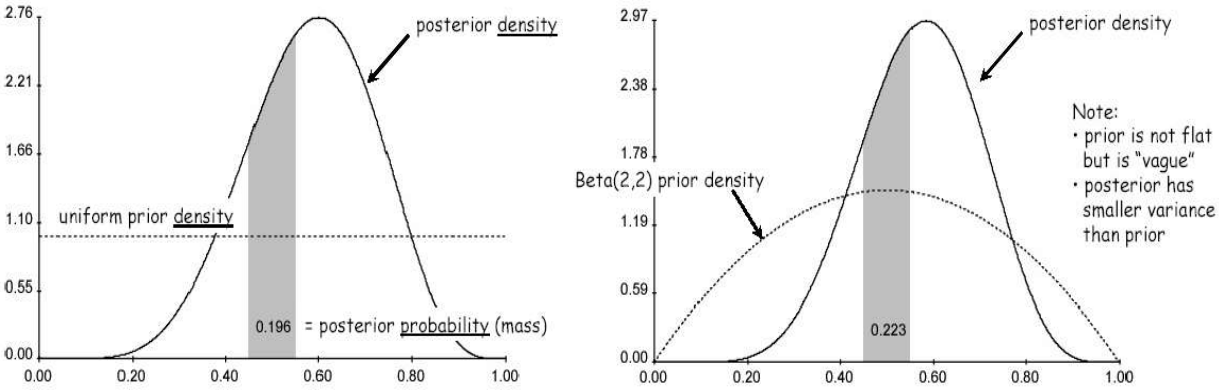
This is the *unconditional probability* of the observed data, which is

$P(\boxdot\boxdot \mid \text{Biased}) \times 0.1 + P(\boxdot\boxdot \mid \text{Fair}) \times 0.9 = 0.03$

Computing this gives that the probability that the die is biased is 0.179.
In this case there is nothing that stops us from doing further investigations by rolling the die again, and we can then use 0.179 as our prior information.

## 4.1 Regarding priors

It is often the case that no prior information exists or that the researcher is unwilling to assign an informative prior, in order not to influence the result with personal opinions. Then an *uninformative* (*flat*) prior can be assigned. It is also possible to assign *vague* priors (Huelsenbeck et al., 2002).




If a flat prior is used the posterior distribution will be proportional to the likelihood.

## 4.2 How to compute - MCMC

The secret behind the increasing popularity of Bayesian analysis lies in the application of Markov Chain Monte Carlo to compute the posterior probability density. A Markov chain is designed, that has as its state space the parameters of the statistical model, and a stationary distribution that is the posterior probability density of the parameters. The chain is then run for a long time (typically millions of steps), and sampled from regularly.

To design the Markov chain properly, it is sufficient to manipulate the transition matrix. How this is done is seen in the algorithm (Metropolis-Hastings):

- Start with random tree and parameters
- In each generation, randomly propose either
  - a new tree topology or
  - a new value for a model parameter
- If the proposed tree has higher posterior probability, $\pi_{\text{proposed}}$, than the current tree, $\pi_{\text{current}}$, the transition is accepted.
- If the proposed tree has lower posterior probability than the current tree, the transition is accepted with probability $\pi_{\text{proposed}}/\pi_{\text{current}}$.
- Every *k* generations, save the current tree.
- After *n* generations, end the Markov chain, and use the saved trees to draw conclusions, using histograms, means, credibility intervals, etc.

These steps ensure that the Markov Chain does indeed have the posterior probability as its stationary distribution. First of all, the transition mechanism assures that the corresponding graph is connected. Furthermore, all nodes of the graph have the same (theoretically infinite) number of neighbors. Along the 'borders' of the graph, this property can be assured by a technique called 'back-reflection'. Thus all nodes have the same degree. Local equilibrium is satisfied. Assume $\pi_i < \pi_j$ :

$$\pi_i P_{i,j} = \frac{\pi_i}{\deg(s_i)} \times \frac{\pi_j}{\pi_i} =$$

Every neighbor of $s_i$ is suggested with probability $1/\deg(s_i)$. The transition is accepted with probability $\pi_j/\pi_i$.

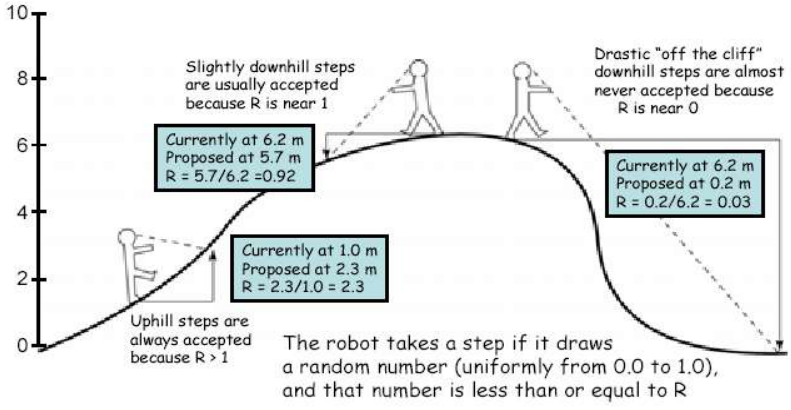
$$\frac{\pi_j}{\deg(s_i)} = \frac{\pi_j}{\deg(s_j)} =$$

All nodes have the same degree.

$$\pi_j P_{j,i}$$

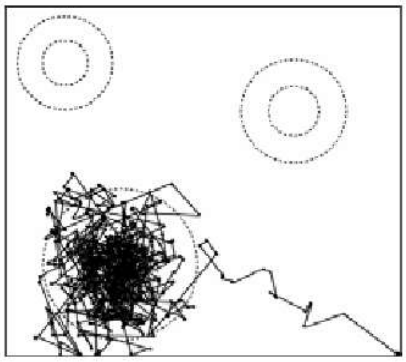
The transition from $\pi_j$ to $\pi_i$ is accepted with probability 1.

This algorithm can be visualized as a robot walking around in a multidimensional landscape. According to its instructions it will consider randomly chosen steps, always accepting up-hill steps and sometimes accepting down-hill steps.




The Markov Chain Monte Carlo algorithm illustrated as a robot walking around in a landscape of posterior probabilities.

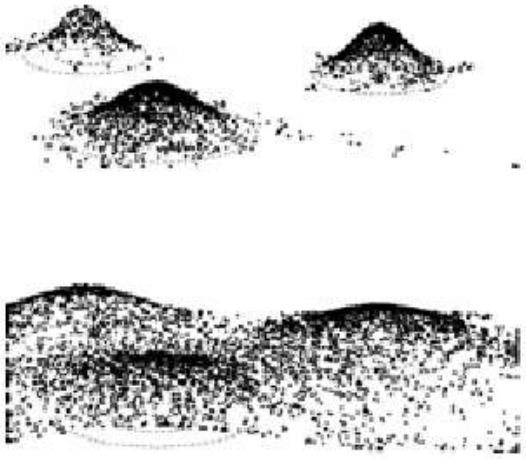
## 4.3Getting stuck on a hill – MCMCMC or MC$^3$



A landscape with several hilltops. The chain gets stuck on the first hill it encounters.

A problem with the MCMC 'robot' is that it has a tendency to get stuck on a hill, once it has encountered it (Larget and Simon, 1999). The above algorithm assures that the stationary distribution of the Markov chain is the posterior probability density for the state space. Therefore the number of visits to a state, $s_i$, will be proportional to to the posterior probability density at this state, $\pi_i$, as the number of steps goes towards infinity, but since we don't have infinite amounts of time to spend, it is necessary to speed things up a bit.

A solution to the problem is Metropolis Coupled MCMC. The idea behind MC$^3$ is simple – instead of running just one Markov chain, several chains are run simultaneously, and all but one of them are *heated*. A heated chain has as its stationary distribution a 'leveled out' version of the posterior probability landscape, in which the hilltops are lower.

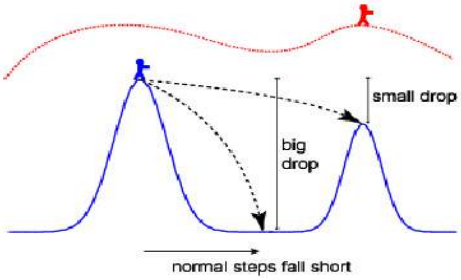


Above is the true probability landscape. Below is the leveled out version experienced by a heated chain.

When several chains are run, the *i*th chain (the cold chain is number 1) will have a stationary distribution proportional to $P(T|D)^{\beta}$, where

$$\beta = \frac{1}{1+(i-1)H}$$

for some heating factor *H*. At every generation in the Markov chain algorithm, a *swap* will be attempted between two randomly chosen chains. If the swap is accepted, the two chains will change status – the colder will become the heated and vice versa. The heated chains will act as *scouts*, more effectively exploring the full



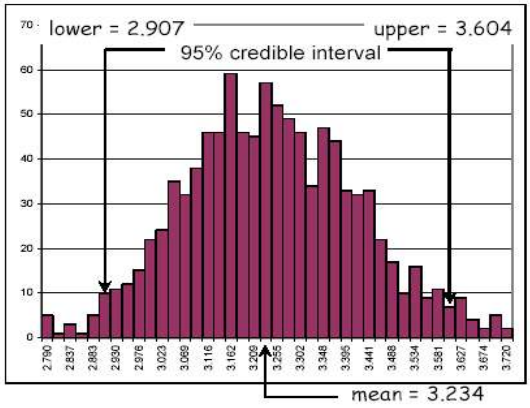

The robot analogy in MCMCMC. The red robot is a heated chain.

breadth of the parameter landscape looking for hilltops. Sampling is done only on the cold chain, since this is the only chain with the correct stationary distribution (Huelsenbeck and Ronquist, 2001).

## 4.4 Drawing conclusions

When the Markov chains have run for long enough, there will be a big set of sampled trees. Depending on the question being asked, conclusions can be drawn in several ways.

- If you want to decide the value of a continuous parameter, this can be achieved using a histogram. The range of the parameter is divided into intervals, and the number of trees in each interval is recorded. Credibility intervals are also easy to compute.



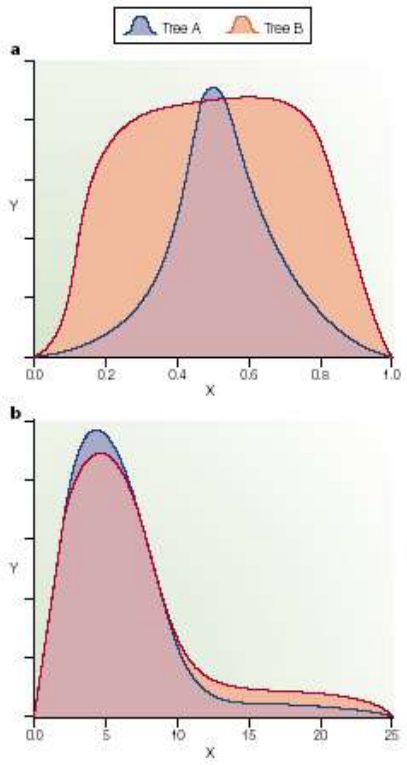

- To determine which tree correctly describes the phylogeny of the studied group of organisms, the common method is to sort all sampled trees according to their posterior probabilities, and then pick the most probable trees, until the cumulative probability is 0.95 (Huelsenbeck et al., 2002).
- If the question is whether a certain group of organisms is monophyletic (if it constitutes a branch in the tree), you just check how many of the sampled trees claim that it is. If it is monophyletic in 74% of the trees, the probability that it is monophyletic is 74%. (Lewis, 2001)

## 4.5 Marginal vs joint estimation

Commonly the prior used in Bayesian analysis of phylogeny is flat. The posterior probabilities will then be proportional to the likelihood. Indeed, the team behind the MrBayes software for performing such analysis claim (Huelsenbeck et al., 2001) that

"*This* [Bayesian inference] *is roughly equivalent to performing a maximum likelihood analysis with bootstrap resampling, but much faster*."

One difference is worth noting, though, and that is the distinction between *marginal* and *joint* estimation. Many of the model parameters might not be of direct interest, but they must be dealt with, since they are part of the likelihood equation. Maximum likelihood aims to find the tallest peak in the parameter landscape (joint estimation), whereas Bayesian inference is measuring the volume under a 'posterior probability surface' (marginal estimation), and will therefore choose a somewhat lower peak if it is wide enough to have a larger volume. Statistically this makes perfect sense. One must bear in mind that the tree suggested by any method is subject to several sources of uncertainty, and it will therefore be an approximation. The output should be interpreted as "This tree, ***or some tree similar to it***, is likely to be the true one." So the likelihood of the suggested tree is not all that matters, the closely related trees are also important (Holder and Lewis, 2003).




Two cases where Maximum likelihood and Bayesian inference will choose differently. On the x axis is a continuous parameter, and on the y axis the likelihood/posterior probability. ML will choose the blue trees, whereas BI will choose the red ones. (Holder and Lewis, 2003)

## 4.5 Unsolved problems

It is still an open question how to determine when a MCMC algorithm has run long enough. Also tests on simulated data sets have revealed some discrepancies between the Maximum likelihood bootstrap estimates and Bayesian posterior probabilities for clades (organisms belonging to the same branch of a tree). In these tests Bayesian posterior probabilities were significantly higher than corresponding bootstrap frequencies for true clades, but Bayesian inference also drew erroneous conclusions more often. The frequency of *type II errors* was lower for the Bayesian analysis, whereas the frequency of *type I errors* was much higher (Erixon et al., 2003; Huelsenbeck et al., 2002).

| | **$H_0$ true** | **$H_0$ false** |
|---|---|---|
| **$H_0$ rejected** | Type I error | Correct |
| **$H_0$ not rejected** | Correct | Type II error |

## 5 References

**Erixon, P et al.** (2003): Reliability of Bayesian Posterior Probabilities and Bootstrap Frequencies in Phylogenetics. *Syst. Biol.* **52**:665-673.

**Holder, M and Lewis, PO** (2003): Phylogeny estimation: traditional and Bayesian approaches. *Nature Reviews Genetics* **4**:275-284.

**Huelsenbeck, JP and Ronquist, F** (2001): MRBAYES: Bayesian inference of phylogenetic trees. *Bioinformatics* **17**:754-755.

**Huelsenbeck, JP et al.** (2001): Bayesian Inference of Phylogeny and Its Impact on Evolutionary Biology. *Science* **294**:2310-2314.

**Huelsenbeck, JP et al.** (2002): Potential Applications and Pitfalls of Bayesian Inference of Phylogeny. *Syst. Biol.* **51**:673-688.

**Larget, B and Simon, DL** (1999): Markov Chain Monte Carlo Algorithms for the Bayesian Analysis of Phylogenetic Trees. *Mol. Biol. Evol.* **16**:750-759.

**Lewis, PO** (2001): Phylogenetic systematics turns over a new leaf. *Trends in Ecology & Evolution* **16**:30-37.