Phylogenetics:
Bayesian Phylogenetic Analysis

COMP 571
Luay Nakhleh, Rice University
Bayes Rule

\[
P(X = x | Y = y) = \frac{P(X = x, Y = y)}{P(Y = y)} = \frac{P(X = x)P(Y = y | X = x)}{\sum_{x'} P(X = x')P(Y = y | X = x')}
\]
Bayes Rule

* Example (from "Machine Learning: A Probabilistic Perspective")

* Consider a woman in her 40s who decides to have a mammogram.

* Question: If the test is positive, what is the probability that she has cancer?

* The answer depends on how reliable the test is!
Bayes Rule

- Suppose the test has a sensitivity of 80%; that is, if a person has cancer, the test will be positive with probability 0.8.

- If we denote by $x=1$ the event that the mammogram is positive, and by $y=1$ the event that the person has breast cancer, then $P(x=1 | y=1) = 0.8$. 
Bayes Rule

* Does the probability that the woman in our example (who tested positive) has cancer equal 0.8?
Bayes Rule

No!

That ignores the prior probability of having breast cancer, which, fortunately, is quite low: \( p(y=1) = 0.004 \)
Further, we need to take into account the fact that the test may be a false positive.

Mammograms have a false positive probability of $p(x=1|y=0)=0.1$. 
Bayes Rule

Combining all these facts using Bayes rule, we get (using $p(y=0)=1-p(y=1)$):

$$p(y = 1|x = 1) = \frac{p(x=1|y=1)p(y=1)}{p(x=1|y=1)p(y=1) + p(x=1|y=0)p(y=0)}$$

$$= \frac{0.8 \times 0.004}{0.8 \times 0.004 + 0.1 \times 0.996}$$

$$= 0.031$$
How does Bayesian reasoning apply to phylogenetic inference?
Assume we are interested in the relationships between human, gorilla, and chimpanzee (with orangutan as an outgroup).

There are clearly three possible relationships.
How does Bayesian reasoning apply to phylogenetic inference? Assume we are interested in the relationships between man, gorilla, and chimpanzee. In the standard case, we need an additional species to root the tree, and the orangutan would be appropriate here. There are three possible ways of arranging these species in a phylogenetic tree: the chimpanzee is our closest relative, the gorilla is our closest relative, or the chimpanzee and the gorilla are each other's closest relatives (Fig. 7.1).

![Bayesian phylogenetic analysis diagram](image)

Fig. 7.1 A Bayesian phylogenetic analysis. We start the analysis by specifying our prior beliefs about the tree. In the absence of background knowledge, we might associate the same probability to each tree topology. We then collect data and use a stochastic evolutionary model and Bayes' theorem to update the prior to a posterior probability distribution. If the data are informative, most of the posterior probability will be focused on one tree (or a small subset of trees in a large tree space).
Before the analysis, we need to specify our prior beliefs about the relationships.

For example, in the absence of background data, a simple solution would be to assign equal probability to the possible trees.
7.2 Bayesian phylogenetic inference

How does Bayesian reasoning apply to phylogenetic inference? Assume we are interested in the relationships between man, gorilla, and chimpanzee. In the standard case, we need an additional species to root the tree, and the orangutan would be appropriate here. There are three possible ways of arranging these species in a phylogenetic tree: the chimpanzee is our closest relative, the gorilla is our closest relative, or the chimpanzee and the gorilla are each other’s closest relatives (Fig. 7.1).

Fig. 7.1 A Bayesian phylogenetic analysis. We start the analysis by specifying our prior beliefs about the tree. In the absence of background knowledge, we might associate the same probability to each tree topology. We then collect data and use a stochastic evolutionary model and Bayes’ theorem to update the prior to a posterior probability distribution. If the data are informative, most of the posterior probability will be focused on one tree (or a small subset of trees in a large tree space).

[This is an uninformative prior]
To update the prior, we need some data, typically in the form of a molecular sequence alignment, and a stochastic model of the process generating the data on the tree.
In principle, Bayes rule is then used to obtain the posterior probability distribution, which is the result of the analysis.

The posterior specifies the probability of each tree given the model, the prior, and the data.
When the data are informative, most of the posterior probability is typically concentrated on one tree (or, a small subset of trees in a large tree space).
7.2 Bayesian phylogenetic inference

How does Bayesian reasoning apply to phylogenetic inference? Assume we are interested in the relationships between man, gorilla, and chimpanzee. In the standard case, we need an additional species to root the tree, and the orangutan would be appropriate here. There are three possible ways of arranging these species in a phylogenetic tree: the chimpanzee is our closest relative, the gorilla is our closest relative, or the chimpanzee and the gorilla are each other’s closest relatives (Fig. 7.1).

![Diagram](image)

1.0 0.0 0.5
Probability

Prior distribution

Data (observations)

Posterior distribution

Fig. 7.1 A Bayesian phylogenetic analysis. We start the analysis by specifying our prior beliefs about the tree. In the absence of background knowledge, we might associate the same probability to each tree topology. We then collect data and use a stochastic evolutionary model and Bayes' theorem to update the prior to a posterior probability distribution. If the data are informative, most of the posterior probability will be focused on one tree (or a small subset of trees in a large tree space).
To describe the analysis mathematically, consider:

- the matrix of aligned sequences $X$
- the tree topology parameter $\tau$
- the branch lengths of the tree $\nu$

(typically, substitution model parameters are also included)

Let $\theta = (\tau, \nu)$
Bayes theorem allows us to derive the posterior distribution as

\[ f(\theta | X) = \frac{f(\theta) f(X | \theta)}{f(X)} \]

where

\[ f(X) = \sum_{\tau} \int_{\nu} f(\nu) f(X | \tau, \nu) \, d\nu \]
Fig. 7.2 Posterior probability distribution for our phylogenetic analysis. The x-axis is an imaginary one-dimensional representation of the parameter space. It falls into three different regions corresponding to the three different topologies. Within each region, a point along the axis corresponds to a particular set of branch lengths on that topology. It is difficult to arrange the space such that optimal branch length combinations for different topologies are close to each other. Therefore, the posterior distribution is multimodal. The area under the curve falling in each tree topology region is the posterior probability of that tree topology.

Even though our model is as simple as phylogenetic models come, it is impossible to portray its parameter space accurately in one dimension. However, imagine for a while that we could do just that. Then the parameter axis might have three distinct regions corresponding to the three different tree topologies (Fig. 7.2). Within each region, the different points on the axis would represent different branch length values. The one-dimensional parameter axis allows us to obtain a picture of the posterior probability function or surface. It would presumably have three distinct peaks, each corresponding to an optimal combination of topology and branch lengths.

To calculate the posterior probability of the topologies, we integrate out the model parameters that are not of interest, the branch lengths in our case. This corresponds to determining the area under the curve in each of the three topology regions. A Bayesian would say that we are marginalizing or deriving the marginal probability distribution on topologies.

Why is it called marginalizing? Imagine that we represent the parameter space in a two-dimensional table instead of along a single axis (Fig. 7.3). The columns in this table might represent different topologies and the rows different branch length values. Since the branch lengths are continuous parameters, there would actually be

The marginal probability distribution on topologies
Why are they called marginal probabilities?

<table>
<thead>
<tr>
<th>Branch length vectors</th>
<th>Topologies</th>
<th>Joint probabilities</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \mathbf{V}^A )</td>
<td>( \tau_A )</td>
<td>0.10 0.07 0.12 0.29</td>
</tr>
<tr>
<td>( \mathbf{V}^B )</td>
<td>( \tau_B )</td>
<td>0.05 0.22 0.06 0.33</td>
</tr>
<tr>
<td>( \mathbf{V}^C )</td>
<td>( \tau_C )</td>
<td>0.05 0.19 0.14 0.38</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.20 0.48 0.32</td>
</tr>
</tbody>
</table>

Marginal probabilities
Markov chain Monte Carlo Sampling
In most cases, it is impossible to derive the posterior probability distribution analytically.

Even worse, we can’t even estimate it by drawing random samples from it.

The reason is that most of the posterior probability is likely to be concentrated in a small part of a vast parameter space.
The solution is to estimate the posterior probability distribution using Markov chain Monte Carlo sampling, or MCMC for short.

- Monte Carlo = random simulation
- Markov chain = the state of the simulator depends only on the current state
Irreducible Markov chains (their topology is strongly connected) have the property that they converge towards an equilibrium state (stationary distribution) regardless of starting point.

We just need to set up a Markov chain that converges onto our posterior probability distribution!
Stationary Distribution of a Markov Chain

\[ P(x_{i+1} = 0 | x_i = 0) = 0.4 \]
\[ P(x_{i+1} = 1 | x_i = 0) = 0.6 \]
\[ P(x_{i+1} = 0 | x_i = 1) = 0.9 \]
\[ P(x_{i+1} = 1 | x_i = 1) = 0.1 \]
Stationary Distribution of a Markov Chain

Consider the simplest possible Markov chain: one with two states. The full probability statements that correspond to the graph are:

$P(x_{i+1} = 0 | x_i = 0) = 0.4$

$P(x_{i+1} = 1 | x_i = 0) = 0.6$

$P(x_{i+1} = 0 | x_i = 1) = 0.9$

$P(x_{i+1} = 1 | x_i = 1) = 0.1$

What are $P(x_i = 0 | x_0 = 0)$, $P(x_i = 1 | x_0 = 0)$, $P(x_i = 0 | x_0 = 1)$, $P(x_i = 1 | x_0 = 1)$?
Stationary Distribution of a Markov Chain

A Markov chain can be defined by describing the full set of probability statements that define the rules for the state of the chain in the next step (step $i+1$) given its current state (in step $i$). These transition probabilities are analogous to transition rates if we are working with continuous-time Markov processes.

Consider the simplest possible Markov chain: one with two states (0, and 1) that operates in discrete time. The figure to the right shows the states in circles. The transition probabilities are shown as arcs connecting the states with the probabilities next to the line.

The full probability statements that correspond to the graph are:

- $P(x_{i+1} = 0 | x_i = 0) = 0.4$
- $P(x_{i+1} = 1 | x_i = 0) = 0.6$
- $P(x_{i+1} = 0 | x_i = 1) = 0.9$
- $P(x_{i+1} = 1 | x_i = 1) = 0.1$

Note that, because these are probabilities, some of them must sum to one. In particular, if we are in a particular state at step $i$, we can call the state $x_i$. In the next step, we must have some state so $\sum_j P(j) = 1$ for every possible $x_i$.

Note that the state at step $i+1$ only depends on the state at step $i$. This is the Markov property. More formally, we could state it as:

$$P(x_{i+1} = j | x_i) = \sum_k P(x_{i+1} = j | x_{i+1} = k, x_i)$$

where $k$ is a positive integer. What this probability statement is saying is that, conditional on $x_i$, the state at $i+1$ is independent of any point before $i$. So when working with Markov chains, we don't need to concern ourselves with the full history of the chain; merely knowing the state at the previous step is enough.

Clearly if we know $x_i$ and the transition probabilities, then we can make a probabilistic statement about the state in the next iteration (in fact the transition probabilities are these probabilistic statements). But we can also think about the probability that the chain will be in a particular state two steps from now:

$$P(x_{i+2} = 0 | x_i = 0) = P(x_{i+2} = 0 | x_{i+1} = 1)P(x_{i+1} = 1 | x_i = 0) + P(x_{i+2} = 0 | x_{i+1} = 0)P(x_{i+1} = 0 | x_i = 0)$$

Here we are exploiting the fact that the same "rules" (transition probabilities) apply when we consider state changes between $i+1$ and $i+2$. If the transition probabilities are fixed throughout the running of the Markov chain, then we are dealing with a time-homogeneous Markov chain.

There are second-order Markov processes that depend on the two previous states, and third-order Markov processes etc. But in this course, we'll just be dealing with the simplest Markov chains which only depend on the current state.
Stationary Distribution of a Markov Chain

A Markov chain can be defined by describing the full set of probability statements that define the rules for the state of the chain in the next step ($x_{i+1}$) given its current state ($x_i$). These transition probabilities are analogous to transition rates if we are working with continuous-time Markov processes.

Consider the simplest possible Markov chain: one with two states (0, and 1) that operates in discrete time. The figure shows the states in circles. The transition probabilities are shown as arcs connecting the states with the probabilities next to the line.

The full probability statements that correspond to the graph are:

- $P(x_{i+1} = 0 | x_i = 0) = 0.4$
- $P(x_{i+1} = 0 | x_i = 1) = 0.9$
- $P(x_{i+1} = 1 | x_i = 0) = 0.6$
- $P(x_{i+1} = 1 | x_i = 1) = 0.1$

Note that, because these are probabilities, some of them must sum to one. In particular, if we are in a particular state at step $i$, we can call the state $x_i$. I n the next step, we must have some state so

- $\sum_j P_j = 1$ for every possible $x_i$.

Note that the state at step $i+1$ only depends on the state at step $i$. This is the Markov property.

More formally we could state it as:

- $P(x_{i+1} = k | x_i, x_{i-1}, \ldots, x_1) = P(x_{i+1} = k | x_i)$

What this probability statement is saying is that, conditional on $x_i$, the state at $i+1$ is independent of the state at any point before $i$. So when working with Markov chains we don't need to concern ourselves with the full history of the chain, merely knowing the state at the previous step is enough.

Clearly if we know $x_i$ and the transition probabilities, then we can make a probabilistic statement about the state in the next iteration (in fact the transition probabilities are these probabilistic statements). But we can also think about the probability that the chain will be in a particular state two steps from now:

- $P(x_{i+2} = 0 | x_i = 0) = P(x_{i+2} = 0 | x_{i+1} = 1)P(x_{i+1} = 1 | x_i = 0) + P(x_{i+2} = 0 | x_{i+1} = 0)P(x_{i+1} = 0 | x_i = 0)$

Here we are exploiting the fact that the same "rules" (transition probabilities) apply when we consider state changes between $i+1$ and $i+2$. If the transition probabilities are fixed through the running of the Markov chain, then we are dealing with a time-homogeneous Markov chain.

There are second-order Markov processes that depend on the two previous states, and third-order Markov process etc. But in this course, we'll just be dealing with the simplest Markov chains which only depend on the current state.
Stationary Distribution of a Markov Chain

Consider the simplest possible Markov chain: one with two states, 0 and 1. The transition probabilities are shown in the diagram:

- From state 0 to state 0: 0.4
- From state 0 to state 1: 0.6
- From state 1 to state 0: 0.1
- From state 1 to state 1: 0.9

The full probability statements that correspond to the graph are:

\[ P(x_{i+1} = 0 | x_i = 0) = 0.4 \]
\[ P(x_{i+1} = 1 | x_i = 0) = 0.6 \]
\[ P(x_{i+1} = 0 | x_i = 1) = 0.1 \]
\[ P(x_{i+1} = 1 | x_i = 1) = 0.9 \]

Note that the state at step \( i + 2 \) only depends on the state at step \( i + 1 \) and \( i \), so the Markov property holds. More formally, we could state it as:

\[ P(x_{i+2} = j | x_i = i, x_{i+1} = k) = P(x_{i+2} = j | x_i = i) \]

This is the Markov property. More formally we could state it as: \( P(x_{i+2} = j | x_i = i, x_{i+1} = k) = P(x_{i+2} = j | x_i = i) \), but in this course, we'll just be dealing with the simplest Markov chains which only depend on the current state.

A Markov chain can be defined by describing the full set of probability statements that define the transitions between states. These rules for the state of the chain in the next step (step \( i + 1 \)) given its current state (in step \( i \)):

- \( P(x_{i+1} = 0 | x_i = 0) = 0.4 \)
- \( P(x_{i+1} = 1 | x_i = 0) = 0.6 \)
- \( P(x_{i+1} = 0 | x_i = 1) = 0.1 \)
- \( P(x_{i+1} = 1 | x_i = 1) = 0.9 \)

The transition probabilities are analogous to transition rates if we are working with continuous-time Markov processes.

Transition probabilities are positive numbers that sum to one:

\[ \sum_j P(x_{i+1} = j | x_i = i) = 1 \]

Note that, because these are probabilities, some of them must sum to one.

The probability stabilizes to a steady state distribution. Knowing whether the chain started in state 0 or 1 tells you very little about the state of the chain in step 15. Technically, the stationary distribution is the probability distribution that is invariant under the transition probabilities, that is, the probability distribution that remains unchanged by the transition operation.

The stationary distribution for this Markov chain is given by:

\[ P(x_{i+1} = 0 | x_i = 0) = 0.3, \quad P(x_{i+1} = 1 | x_i = 0) = 0.7 \]
\[ P(x_{i+1} = 0 | x_i = 1) = 0.9, \quad P(x_{i+1} = 1 | x_i = 1) = 0.1 \]

The full probability statements that correspond to the graph are:

\[ P(x_{i+2} = 0 | x_i = 0, x_{i+1} = 0) = 0.4 \]
\[ P(x_{i+2} = 0 | x_i = 0, x_{i+1} = 1) = 0.6 \]
\[ P(x_{i+2} = 0 | x_i = 1, x_{i+1} = 0) = 0.1 \]
\[ P(x_{i+2} = 0 | x_i = 1, x_{i+1} = 1) = 0.9 \]

The stationary distribution is found by solving the system of equations:

\[ \sum_j P(x_{i+1} = j | x_i = i) = 1 \]

For this Markov chain, the stationary distribution is:

\[ P(x_i = 0) = 0.3, \quad P(x_i = 1) = 0.7 \]

These are the long-run proportions of time spent in each state.

The transition matrix for this Markov chain is:

\[ P = \begin{pmatrix}
0.4 & 0.6 \\
0.1 & 0.9
\end{pmatrix} \]

The stationary distribution is given by the eigenvector of the transition matrix corresponding to the eigenvalue 1.
Stationary Distribution of a Markov Chain

![Diagram of a two-state Markov chain with transition probabilities 0.4 to 0, 0.6 from 0 to 1, 0.9 from 1 to 0, and 0.1 from 1 to 1.]

- The same probability regardless of starting state!

<table>
<thead>
<tr>
<th>State</th>
<th>Transition to State</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0.4 → 0, 0.6 → 1</td>
</tr>
<tr>
<td>1</td>
<td>0.9 → 0, 0.1 → 1</td>
</tr>
</tbody>
</table>

- Stationary Distribution

\[ P(x_i = 0 | x_0 = 0) \]

\[ P(x_i = 0 | x_0 = 1) \]
A Markov chain can be defined by describing the full set of probability statements that define the rules for the state of the chain in the next step (step $i+1$). This is the Markov property. More formally we could state it as:

$$P(x_i = j | x_{i-1} = k) = P(x_i = j | x_{i-1}) = P(x_{i+1} = j | x_i = k)$$

Note that the state at step $i+1$ is independent on the state at any point before $i$. So when working with Markov processes, the state at the previous step is enough.

Consider the simplest possible Markov chain: one with two states, 0 and 1. In particular, if we are in a particular state at step $i$, then at step $i+1$ the state of the chain can change to one of two possible states, 0 or 1, with some probability. The probability of staying at the same state remains the same at every step:

$$P(x_i = j | x_{i+1} = k) = P(x_i = j | x_{i}) = P(x_{i+1} = j | x_i = k)$$

Clearly if we know the state at the previous step is enough.

If we consider the state at step 1, then for every possible state $x$, we have:

$$P(x_1 = j | x_0 = k) = P(x_1 = j | x_0) = P(x_2 = j | x_1)$$

This is the Markov property. More formally we could state it as:

$$P(x_i = j | x_{i-1}) = P(x_i = j)$$

Note that the state at step 1 is independent on the state at any point before $i$.

Examples of Markov chains include the weather, the stock market, the probability of nuclear war, and other phenomena where the only concern is the probability of going on to the next state. Consider the simplest possible Markov chain: one with two states, 0 and 1. If the transition probabilities are fixed through the iteration of a two-state Markov chain, there is a stationary distribution. If the transition probabilities are fixed through the iteration of a two-state Markov chain, there is a stationary distribution. Knowing whether the chain converges to this distribution is important. If we consider an even larger number of iterations (e.g. the state at step 100), then the probabilities of being in state 0 as a function of the step number, $i$, stabilize to a steady state distribution. Knowing whether the chain converges to this distribution is important.

So when working with Markov processes, the state at the previous step is enough. Therefore, if we know the state at step 1, then for every possible state $x$, we have:

$$P(x_i = j | x_{i-1} = k) = P(x_i = j | x_{i}) = P(x_{i+1} = j | x_i = k)$$

Note that the state at step $i+1$ is independent on the state at any point before $i$. So when working with Markov processes, the state at the previous step is enough.

Consider the simplest possible Markov chain: one with two states, 0 and 1. In particular, if we are in a particular state at step $i$, then at step $i+1$ the state of the chain can change to one of two possible states, 0 or 1, with some probability. The probability of staying at the same state remains the same at every step:

$$P(x_i = j | x_{i+1} = k) = P(x_i = j | x_{i}) = P(x_{i+1} = j | x_i = k)$$

Clearly if we know the state at the previous step is enough.

If we consider the state at step 1, then for every possible state $x$, we have:

$$P(x_1 = j | x_0 = k) = P(x_1 = j | x_0) = P(x_2 = j | x_1)$$

This is the Markov property. More formally we could state it as:

$$P(x_i = j | x_{i-1}) = P(x_i = j)$$

Note that the state at step 1 is independent on the state at any point before $i$.

Examples of Markov chains include the weather, the stock market, the probability of nuclear war, and other phenomena where the only concern is the probability of going on to the next state. Consider the simplest possible Markov chain: one with two states, 0 and 1. If the transition probabilities are fixed through the iteration of a two-state Markov chain, there is a stationary distribution. If the transition probabilities are fixed through the iteration of a two-state Markov chain, there is a stationary distribution. Knowing whether the chain converges to this distribution is important. If we consider an even larger number of iterations (e.g. the state at step 100), then the probabilities of being in state 0 as a function of the step number, $i$, stabilize to a steady state distribution. Knowing whether the chain converges to this distribution is important.

So when working with Markov processes, the state at the previous step is enough. Therefore, if we know the state at step 1, then for every possible state $x$, we have:

$$P(x_i = j | x_{i-1} = k) = P(x_i = j | x_{i}) = P(x_{i+1} = j | x_i = k)$$

Note that the state at step $i+1$ is independent on the state at any point before $i$. So when working with Markov processes, the state at the previous step is enough.

Consider the simplest possible Markov chain: one with two states, 0 and 1. In particular, if we are in a particular state at step $i$, then at step $i+1$ the state of the chain can change to one of two possible states, 0 or 1, with some probability. The probability of staying at the same state remains the same at every step:

$$P(x_i = j | x_{i+1} = k) = P(x_i = j | x_{i}) = P(x_{i+1} = j | x_i = k)$$

Clearly if we know the state at the previous step is enough.

If we consider an even larger number of iterations (e.g. the state at step 100), then the probabilities of being in state 0 as a function of the step number, $i$, stabilize to a steady state distribution. Knowing whether the chain converges to this distribution is important.
Stationary Distribution of a Markov Chain

where does the 0.6 come from?
Stationary Distribution of a Markov Chain

The stationary distribution of a Markov chain is a probability distribution that is invariant under the transition probabilities of the chain. In other words, if the chain is in the stationary distribution at some step, it will remain in that distribution for all subsequent steps.

For a two-state Markov chain with transition probabilities of 0.4 from state 0 to state 1 and 0.6 from state 1 to state 0, the stationary distribution can be found by solving the system of equations:

\[
\begin{align*}
\pi_0 &= 0.4\pi_1 + 0.6\pi_0 \\
\pi_1 &= 0.6\pi_0 + 0.4\pi_1
\end{align*}
\]

Given that \(\pi_0 + \pi_1 = 1\), we can solve this system to find the stationary distribution \(\pi_0 = 0.6\) and \(\pi_1 = 0.4\).

Where does the 0.6 come from?

The value of 0.6 in the stationary distribution represents the probability of being in state 0 in the limiting behavior of the Markov chain. This means that, in the long run, the proportion of time spent in state 0 is 60%.

Stationary distribution: \(\pi_0 = 0.6\), \(\pi_1 = 0.4\)
Stationary Distribution of a Markov Chain

A Markov chain can be defined by describing the full set of probability statements that define the rules for the state of the chain in the next step \((i+1)\) given its current state \((i)\). These transition probabilities are analogous to transition rates if we are working with continuous-time Markov processes.

Consider the simplest possible Markov chain: one with two states (0, and 1) that operates in discrete time. The figure to the right shows the states in circles. The transition probabilities are shown as arcs connecting the states with the probabilities next to the line. The full probability statements that correspond to the graph are:

\[
P(x_{i+1} = 0 | x_i = 0) = 0.4
\]

\[
P(x_{i+1} = 1 | x_i = 0) = 0.6
\]

\[
P(x_{i+1} = 0 | x_i = 1) = 0.9
\]

\[
P(x_{i+1} = 1 | x_i = 1) = 0.1
\]

Note that, because these are probabilities, some of them must sum to one. In particular, if we are in a particular state at step \(i\) we can call the state \(x_i\). In the next step, we have some state so

\[P_j P(x_{i+1} = j | x_i)\]

for every possible \(x_i\).

Note that the state at step \(i+1\) only depends on the state at step \(i\). This is the Markov property. More formally we could state it as:

\[P(x_{i+1} | x_i) = P(x_{i+1} | x_i, x_k)\]

where \(k\) is positive integer. What this probability statement is saying is that, conditional on \(x_i\), the state at \(i+1\) is independent on the state at any point before \(i+1\). So when working with Markov chains we don't need to concern ourselves with the full history of the chain, merely knowing the state at the previous step is enough.

Clearly if we know \(x_i\) and the transition probabilities, then we can make a probabilistic statement about the state in the next iteration (in fact the transition probabilities are these probabilistic statements). But we can also think about the probability that the chain will be in a particular state two steps from now:

\[
P(x_{i+2} = 0 | x_i = 0) = P(x_{i+2} = 0 | x_{i+1} = 1) P(x_{i+1} = 1 | x_i = 0) + P(x_{i+2} = 0 | x_{i+1} = 0) P(x_{i+1} = 0 | x_i = 0)
\]

Here we are exploiting the fact that the same “rules” (transition probabilities) apply when we consider state changes between \(i+1\) and \(i+2\). If the transition probabilities are fixed through the running of the Markov chain, then we are dealing with a time-homogeneous Markov chain. There are second-order Markov processes that depend on the two previous states, and third-order Markov processes etc. But in this course, we'll just be dealing with the simplest Markov chains which only depend on the current state.
A Markov chain can be defined by describing the full set of probability statements that define the rules for the state of the chain in the next step (step \( i + 1 \)) given its current state (in step \( i \)). These transition probabilities are analogous to transition rates if we are working with continuous-time Markov processes.

Consider the simplest possible Markov chain: one with two states (0, and 1) that operates in discrete time. The figure to the right shows the states in circles. The transition probabilities are shown as arcs connecting the states with the probabilities next to the line. The full probability statements that correspond to the graph are:

- \( P(x_{i+1} = 0 | x_i = 0) = 0.4 \)
- \( P(x_{i+1} = 1 | x_i = 0) = 0.6 \)
- \( P(x_{i+1} = 0 | x_i = 1) = 0.9 \)
- \( P(x_{i+1} = 1 | x_i = 1) = 0.1 \)

Note that, because these are probabilities, some of them must sum to one. In particular, if we are in a particular state at step \( i \), we can call the state \( x_i \). I n the next step, we must have some state \( x_{i+1} \) such that \( P_j \) for every possible \( x_i \).

Note that the state at step \( i + 1 \) only depends on the state at step \( i \). This is the Markov property. More formally we could state it as:

\[ P(x_{i+1} | x_i) = P(x_{i+1} | x_i, x_k) \]

where \( k \) is a positive integer. What this probability statement is saying is that, conditional on \( x_i \), the state at \( i + 1 \) is independent on the state at any point before \( i + 1 \). So when working with Markov chains we don't need to concern ourselves with the full history of the chain, merely knowing the state at the previous step is enough.

Clearly if we know \( x_i \) and the transition probabilities, then we can make a probabilistic statement about the state in the next iteration (in fact the transition probabilities are these probabilistic statements). But we can also think about the probability that the chain will be in a particular state two steps from now:

\[ P(x_{i+2} = 0 | x_i = 0) = P(x_{i+2} = 0 | x_{i+1} = 1)P(x_{i+1} = 1 | x_i = 0) + P(x_{i+2} = 0 | x_{i+1} = 0) \]

Imagine infinitely many chains. At equilibrium (steady-state), the “flux out” of each state must be equal to the “flux into” that state.
Imagine infinitely many chains. At equilibrium (steady-state), the “flux out” of each state must be equal to the “flux into” that state.

\[
\pi_0 \mathbb{P}(x_{i+1} = 1 | x_i = 0) = \pi_1 \mathbb{P}(x_{i+1} = 0 | x_i = 1) \\
\frac{\pi_0}{\pi_1} = \frac{\mathbb{P}(x_{i+1} = 0 | x_i = 1)}{\mathbb{P}(x_{i+1} = 1 | x_i = 0)}
\]
Imagine infinitely many chains. At equilibrium (steady-state), the “flux out” of each state must be equal to the “flux into” that state.

\[
\begin{align*}
\pi_0 \mathbb{P}(x_{i+1} = 1 | x_i = 0) & = \pi_1 \mathbb{P}(x_{i+1} = 0 | x_i = 1) \\
\pi_0 \mathbb{P}(x_{i+1} = 0 | x_i = 1) & = \pi_1 \mathbb{P}(x_{i+1} = 1 | x_i = 0)
\end{align*}
\]

\[
\pi_0 = \mathbb{P}(x_i = 0) \quad \pi_1 = \mathbb{P}(x_i = 1)
\]
Stationary Distribution of a Markov Chain

Consider the simplest possible Markov chain: one with two states (0, and 1) that operates in discrete time. The figure to the right shows the states in circles. The transition probabilities are shown as arcs connecting the states with the probabilities next to the line. The full probability statements that correspond to the graph are:

\[
P(x_{i+1} = 0 | x_i = 0) = 0.4\]
\[
P(x_{i+1} = 1 | x_i = 0) = 0.6\]
\[
P(x_{i+1} = 0 | x_i = 1) = 0.9\]
\[
P(x_{i+1} = 1 | x_i = 1) = 0.1\]

Note that the state at step \(i\) only depends on the state at step \(i\). This is the Markov property. More formally we could state it as:

\[
P(x_{i+1} | x_i) = P(x_{i+1} | x_i, x_{i-1}, \ldots, x_0)\]

where \(k\) is a positive integer. What this probability statement is saying is that, conditional on \(x_i\), the state at \(i+1\) is independent of the state at any point before \(i\). So when working with Markov chains we don't need to concern ourselves with the full history of the chain, merely knowing the state at the previous step is enough.

Clearly if we know \(x_i\) and the transition probabilities, then we can make a probabilistic statement about the state in the next iteration (in fact the transition probabilities are these probabilistic statements). But we can also think about the probability that the chain will be in a particular state two steps from now:

\[
P(x_{i+2} = 0 | x_i = 0) = P(x_{i+2} = 0 | x_{i+1} = 1)P(x_{i+1} = 1 | x_i = 0) + P(x_{i+2} = 0 | x_{i+1} = 0)P(x_{i+1} = 0 | x_i = 0)\]

Here we are exploiting the fact that the same “rules” (transition probabilities) apply when we consider state changes between \(i+1\) and \(i+2\). If the transition probabilities are fixed through the running of the Markov chain, then we are dealing with a time-homogeneous Markov chain.

There are second-order Markov processes that depend on the two previous states, and third-order Markov processes etc. But in this course, we’ll just be dealing with the simplest Markov chains which only depend on the current state.

2
State.

Consider state changes between

P statements). But we can also think about the probability that the chain will be in a particular state in the next iteration (in fact the transition probabilities clearly if we know the state at the previous step is enough.

Chains we don’t need to concern ourselves with the full history of the chain, merely knowing the state at

(k = 0) = 0)

Note that the state at step

1

Note that, because these are probabilities, some of them must sum to 1.

This is the Markov property. More formally we could state it as:

If we force the “flux out” and “flux in” to be identical:

At the equilibrium process. This is the joint probability of being in state 0 at step i

How can we show that this is the steady-state (equilibrium) distribution? The flux “out” of each state must be equal to the “flux” into that state for a system to be at equilibrium. Here it is helpful to think about running infinitely many chains. If we have an equilibrium probability distribution,

Steady-state probability distribution (frequently denoted π) is a probability distribution on the state space such that the state at the next step is independent of the current state. That is,

Stationary distribution

After a long enough walk, a Markov chain in which all of the states are connected will be very close to being independent of each other (as Figure shows). Thus, when the chain is running in a time-homogeneous Markov chain.

The stationary distribution is the limiting distribution of the Markov chain. It is the distribution that the chain approaches as the number of iterations increases.

Consider a two-state Markov chain with transition probabilities as shown in the figure.

Stationary Distribution

of a Markov Chain

π0

π1

= P(x_{i+1} = 0|x_i = 1)
P(x_{i+1} = 1|x_i = 0)
Stationary Distribution of a Markov Chain

\[ \frac{\pi_0}{\pi_1} = \frac{\mathbb{P}(x_{i+1} = 0|x_i = 1)}{\mathbb{P}(x_{i+1} = 1|x_i = 0)} \]

\[ \pi_0 + \pi_1 = 1 \]
Stationary Distribution of a Markov Chain

\[
\begin{align*}
\pi_0 &= \frac{\mathbb{P}(x_{i+1} = 0|x_i = 1)}{\mathbb{P}(x_{i+1} = 1|x_i = 0)} \\
\pi_0 + \pi_1 &= 1 \\
\frac{\pi_0}{\pi_1} &= \frac{0.9}{0.6} = 1.5 \\
\pi_0 &= 1.5\pi_1 \\
1.5\pi_1 + \pi_1 &= 1.0 \\
\pi_1 &= 0.4 \\
\pi_0 &= 0.6
\end{align*}
\]
If we can choose the transition probabilities of the Markov chain, then we can construct a sampler that will converge to any distribution that we desire!
Stationary Distribution of a Markov Chain

* For the general case of more than 2 states:

\[
\text{flux out of } j = \pi_j \mathbb{P}(x_{i+1} \in S \neq j | x_i = j) \\
= \pi_j [1 - \mathbb{P}(x_{i+1} \in j | x_i = j)]
\]

\[
\text{flux into } j = \sum_{k \in S \neq j} \pi_k \mathbb{P}(x_{i+1} = j | x_i = k)
\]

\[
\pi_j [1 - \mathbb{P}(x_{i+1} = j | x_i = j)] = \sum_{k \in S \neq j} \pi_k \mathbb{P}(x_{i+1} = j | x_i = k)
\]

\[
\pi_j = \pi_j \mathbb{P}(x_{i+1} = j | x_i = j) + \sum_{k \in S \neq j} \pi_k \mathbb{P}(x_{i+1} = j | x_i = k)
\]

\[
= \sum_{k \in S} \pi_k \mathbb{P}(x_{i+1} = j | x_i = k)
\]


Mixing

* While setting the transition probabilities to specific values affects the stationary distribution, the transition probabilities cannot be determined uniquely from the stationary distribution.
In MCMC we will design a Markov chain such that its stationary distribution will be identical to the posterior. But even a slowly mixing chain will (in theory) eventually be capable of providing transition probabilities, so that our MCMC approximation will quickly converge to the rate of convergence of a chain to its stationary distribution is an aspect of a Markov process.

Thus, that the rate of convergence of a chain to its stationary distribution is an aspect of a Markov process.

Figure 2: The probability of being in state 0 as a function of the step number, for two different starting states (0 and 1) for the Markov process depicted in Figure 1 = 0

Figure 3: The probability of being in state 0 as a function of the step number, for code. Figure depicts the same probability for a process in which the state-changing rates must be to achieve that distribution?

There is a one-to-one mapping between transition probabilities and a stationary distribution. In the form of a question: if we were given the stationary distribution could we decide what the transition rates are so close that they are indistinguishable.

If we consider an even larger number of iterations (e.g. the state at step 100), then the probabilities that the rate of convergence of a chain to its stationary distribution is an aspect of a Markov process.

But clearly these two probabilities are very close to being equal.

How would such a chain behave? The primary di

dropping the “flux out” of each state by a factor of 10. Because the “self-transition” rate would be

Note that the probability stabilizes to a steady state distribution. Knowing whether the chain

starting states (0 and 1) for the Markov process depicted in Figure (20

0.06 and 0.9). Imagine

40)

Note that:

\[ P(x_{i+1} = 1 | x_i = 0) = 0.6 \]

\[ P(x_{i+1} = 0 | x_i = 1) = 0.9 \]

\[ P(x_{i+1} = 1 | x_i = 0) = 0.06 \]

\[ P(x_{i+1} = 0 | x_i = 1) = 0.09 \]

So, if we sum over all possible states at
In MCMC we will design a Markov chain such that its stationary distribution will be identical to the posterior probability distribution over the space of parameters. We will try to design chains in MCMC that are separate from what the stationary distribution is.

Thus, that the rate of convergence of a chain to its stationary distribution is an aspect of a Markov chain that is. But even a slowly mixing chain will (in theory) eventually be capable of providing a posterior approximation that has high transition probabilities, so that our MCMC approximation will quickly converge to the posterior probability distribution over the space of parameters.

In Figure 2, we start with the state-changing rates of the Markov process depicted in Figure 1, dropping the "flux out" of each state by a factor of 10. Because the "self-transition" rate would increase by the appropriate amount, we can still end up in the same stationary distribution. It turns out that the answer is "No." Which can be seen if we examine equation (6) but for a process in which the transition probabilities are scaled down by a factor of 10. Compare this to Figure (3) showing the effect of changing the transition probabilities.

Figure 3: The probability of being in state 0 as a function of the step number, for two different starting states (0 and 1) for the Markov process depicted in Figure (1).

How would such a chain behave? The primary difference is that it would "mix" more slowly. Adjacent steps would be more likely to be in the same state, and it would take a larger number of iterations before the chain "forgets" its starting state. Figure (4) shows the probabilities of starting states (0 and 1) for the Markov process being in state 0 as a function of the step number for the two possible starting states.

It gets tedious to continue this for a large number of steps into the future. But we can ask a form of a question: if we were given the stationary distribution could we decide what the transition rates must be to achieve that distribution?

We just saw how changing transition probabilities will affect the stationary distribution. But clearly these two probabilities are very close to being equal.

Note that the probability stabilizes to a steady state distribution. Knowing whether the chain started in state 0 or 1 tells you very little about the state of the chain in step 15. Technically, the probability of the state at step 15 does depend on the starting state:

\[
\Pr(x_i=0) = \begin{cases} 
0.06 & \text{if } x_i = 0 \\
0.09 & \text{if } x_i = 1
\end{cases}
\]

Note that the probability stabilizes to a steady state distribution. Knowing whether the chain started in state 0 or 1 tells you very little about the state of the chain in step 15. Technically, the probability of the state at step 15 does depend on the starting state:

\[
\Pr(x_i=0) = \begin{cases} 
0.06 & \text{if } x_i = 0 \\
0.09 & \text{if } x_i = 1
\end{cases}
\]

Note that:

\[
\Pr(x_{i+1} = 1 | x_i = 0) = 0.6 \quad \Pr(x_{i+1} = 0 | x_i = 1) = 0.9
\]

\[
\Pr(x_{i+1} = 1 | x_i = 0) = 0.06 \quad \Pr(x_{i+1} = 0 | x_i = 1) = 0.09
\]
Setting the transition probabilities to lower values resulted in a chain that “mixed” more slowly: Adjacent steps would be more likely to be in the same state and, thus, would require a larger number of iterations before the chain “forgets” its starting state.
Mixing

- The rate of convergence of a chain to its stationary distribution is an aspect of a Markov chain that is separate from what the stationary distribution is.
Mixing

* In MCMC, we will design a Markov chain whose stationary distribution is identical to the posterior probability distribution over the space of parameters.

* We try to design chains that have high transition probabilities to achieve faster convergence.
Detailed Balance

- In practice, the number of states is very large.
- Setting the transition probabilities so that we have equal flux into and out of any state is tricky.
- What we use instead is detailed balance.
Detailed Balance

* We restrict ourselves to Markov chains that satisfy detailed balance for all pairs of states $j$ and $k$:

$$\pi_j P(x_{i+1} = k | x_i = j) = \pi_k P(x_{i+1} = j | x_i = k)$$

(equivalently: $\frac{\pi_j}{\pi_k} = \frac{P(x_{i+1} = j | x_i = k)}{P(x_{i+1} = k | x_i = j)}$)
This can be achieved using several different methods, the most flexible of which is known as the **Metropolis algorithm** and its extension, the **Metropolis-Hastings method**.
In the Metropolis-Hastings algorithm, we choose rules for constructing a random walk through the parameter space.

We adopt transition probabilities such that the stationary distribution of our Markov chain is equal to the posterior probability distribution:

\[ \pi_{\theta_j} = \mathbb{P}(\theta_j | \text{Data}) \]
\[
\frac{t_{\ell,k}}{t_{k,\ell}} = \frac{\pi_k}{\pi_\ell} = \left(\frac{\mathbb{P}(D|\theta_j = k)\mathbb{P}(\theta_j = k)}{\mathbb{P}(D)}\right) / \left(\frac{\mathbb{P}(D|\theta_j = \ell)\mathbb{P}(\theta_j = \ell)}{\mathbb{P}(D)}\right)
\]
\[
\frac{t_{\ell,k}}{t_{k,\ell}} = \frac{\pi_k}{\pi_{\ell}} = \left( \frac{\mathbb{P}(D|\theta_j = k)\mathbb{P}(\theta_j = k)}{\mathbb{P}(D)} \right) / \left( \frac{\mathbb{P}(D|\theta_j = \ell)\mathbb{P}(\theta_j = \ell)}{\mathbb{P}(D)} \right)
\]

desired property
\[
\frac{t_{\ell,k}}{t_{k,\ell}} = \frac{\pi_k}{\pi_\ell} = \frac{\binom{\mathbb{P}(D|\theta_j = k)\mathbb{P}(\theta_j = k)}{\mathbb{P}(D)}}{\binom{\mathbb{P}(D|\theta_j = \ell)\mathbb{P}(\theta_j = \ell)}{\mathbb{P}(D)}}
\]
\[
\frac{t_{\ell,k}}{t_{k,\ell}} = \frac{\pi_k}{\pi_\ell} = \left( \frac{\mathbb{P}(D | \theta_j = k) \mathbb{P}(\theta_j = k)}{\mathbb{P}(D)} \right) \Bigg/ \left( \frac{\mathbb{P}(D | \theta_j = \ell) \mathbb{P}(\theta_j = \ell)}{\mathbb{P}(D)} \right)
\]

desired property

detailed balance

\(\mathbb{P}(D)\) cancels out, so doesn't need to be computed!
Therefore, we need to set the transition probabilities so that

\[ \frac{t_{\ell,k}}{t_{k,\ell}} = \frac{\mathbb{P}(D|\theta_j = k)\mathbb{P}(\theta_j = k)}{\mathbb{P}(D|\theta_j = \ell)\mathbb{P}(\theta_j = \ell)} \]
However, an important problem arises when doing this, which can be illustrated as follows:

- when dealing with states $k, l$, it could be that we need to set $t_{k, l} = 1$ and $t_{l, k} = 0.5$
- when dealing with states $k, m$, it could be that we need to set $t_{k, m} = 0.3$ and $t_{m, k} = 0.1$
- Then, we have $t_{k, m} + t_{k, l} = 1.3$, which violates the fundamental rules of probability!
**Solution:**

* view the transition probability as a joint event: (1) the move is proposed with probability $q$, and (2) the move is accepted with probability $\alpha$.

* If we denote by $x'_{i+1}$ the state proposed at step $i+1$, then

$$q(j, k) = \mathbb{P}(x'_{i+1} = k | x_i = j)$$

$$\alpha(j, k) = \mathbb{P}(x_{i+1} = k | x'_i = j, x'_{i+1} = k)$$
We can choose proposal probabilities that sum to one for all the state-changing transitions.

Then, we can multiply them by the appropriate acceptance probabilities (keeping them as high as possible, but ≤ 1).

We get

\[ \frac{t_{\ell,k}}{t_{k,\ell}} = \frac{q(\ell, k)\alpha(\ell, k)}{q(k, \ell)\alpha(k, \ell)} = \frac{\mathbb{P}(D|\theta_j = k)\mathbb{P}(\theta_j = k)}{\mathbb{P}(D|\theta_j = \ell)\mathbb{P}(\theta_j = \ell)} \]
We have flexibility in selecting how we perform proposals on new states in MCMC.

We have to ensure that \( q(l,k) > 0 \) whenever \( q(k,l) > 0 \) (it is fine if both are 0, but we can't have one being 0 and the other greater than 0).
However, once we have chosen a proposal scheme, we do not have much flexibility in choosing whether or not to accept a proposal.

\[
\frac{t_{\ell,k}}{t_{k,\ell}} = \frac{q(\ell, k) \alpha(\ell, k)}{q(k, \ell) \alpha(k, \ell)} = \frac{\mathbb{P}(D|\theta_j = k) \mathbb{P}(\theta_j = k)}{\mathbb{P}(D|\theta_j = \ell) \mathbb{P}(\theta_j = \ell)}
\]
\[
\frac{t_{\ell, k}}{t_{k, \ell}} = \frac{q(\ell, k) \alpha(\ell, k)}{q(k, \ell) \alpha(k, \ell)} = \frac{\mathbb{P}(D|\theta_j = k)\mathbb{P}(\theta_j = k)}{\mathbb{P}(D|\theta_j = \ell)\mathbb{P}(\theta_j = \ell)}
\]
\[
\frac{t_{\ell,k}}{t_{k,\ell}} = \frac{q(\ell, k)\alpha(\ell, k)}{q(k, \ell)\alpha(k, \ell)} = \frac{\mathbb{P}(D|\theta_j = k)\mathbb{P}(\theta_j = k)}{\mathbb{P}(D|\theta_j = \ell)\mathbb{P}(\theta_j = \ell)}
\]

\[
\frac{\alpha(\ell, k)}{\alpha(k, \ell)} = \frac{\mathbb{P}(D|\theta_j = k)\mathbb{P}(\theta_j = k)q(k, \ell)}{\mathbb{P}(D|\theta_j = \ell)\mathbb{P}(\theta_j = \ell)q(\ell, k)}
\]
\[
\frac{t_{\ell,k}}{t_{k,\ell}} = \frac{q(\ell, k)\alpha(\ell, k)}{q(k, \ell)\alpha(k, \ell)} = \frac{\mathbb{P}(D|\theta_j = k)\mathbb{P}(\theta_j = k)}{\mathbb{P}(D|\theta_j = \ell)\mathbb{P}(\theta_j = \ell)}
\]

\[
\frac{\alpha(\ell, k)}{\alpha(k, \ell)} = \frac{\mathbb{P}(D|\theta_j = k)\mathbb{P}(\theta_j = k)q(k, \ell)}{\mathbb{P}(D|\theta_j = \ell)\mathbb{P}(\theta_j = \ell)q(\ell, k)}
\]

acceptance ratio = \(\text{(likelihood ratio)}\) \(\text{(prior ratio)}\) \(\text{(Hastings ratio)}\)
The central idea is to make small random changes to some current parameter values, and then accept or reject the changes according to the appropriate probabilities.
Bayesian phylogenetic analysis using Markov chain Monte Carlo (MCMC) theory.

1. Start at an arbitrary point ($q$).
2. Make a small random move (to $q'$).
   - If $r > 1$: new state accepted.
   - If $r < 1$: new state accepted with probability $r$.

3. Calculate height ratio ($r$) of new state ($q'$) to old state ($q$) and always accept.

4. Go to step 2.

Markov chain Monte Carlo steps:

- Topology A: 20% posterior probability.
- Topology B: 48% posterior probability.
- Topology C: 32% posterior probability.

The three ratios in the last equation are referred to as the prior ratio, the likelihood ratio, and the proposal ratio (or Hastings ratio), respectively. The first two ratios correspond to the ratio of the numerators in Bayes' theorem; note that the complex...
Markov chain Monte Carlo steps

1. Start at an arbitrary point ($\theta$)
2. Make a small random move (to $\theta^*$)
3. Calculate height ratio ($r$) of new state (to $\theta^*$) to old state ($\theta$)
   (a) $r > 1$: new state accepted
   (b) $r < 1$: new state accepted with probability $r$
      if new state rejected, stay in old state
4. Go to step 2
The three ratios in the last equation are referred to as the prior ratio, the likelihood ratio, and the proposal ratio (or Hastings ratio), respectively. The first two ratios correspond to the ratio of the numerators in Bayes' theorem; note that the complex
An example of a proposal mechanism is the \textit{beta proposal}:

\begin{itemize}
  \item Assume the current values are \((x_1, x_2)\);
  \item Multiply them with a value \(\alpha\);
  \item Pick new values from \text{Beta}(\alpha x_1, \alpha x_2);
  \item Pick new values from \text{Beta}(\alpha x_1, \alpha x_2)
\end{itemize}
A simpler proposal mechanism is to define a continuous uniform distribution of width $w$, centered on the current value $x$, and the new value $x^*$ is drawn from this distribution.
More complex moves are needed to change tree topology.

A common type uses operations such as SPR, TBR, and NNI.
Burn-in, mixing, and convergence
If the chain is started from a random tree and arbitrarily chosen branch lengths, chances are that the initial likelihood is low.

The early phase of the run in which the likelihood increases very rapidly towards regions in the posterior with high probability mass is known as the burn-in.
7.4 Burn-in, mixing and convergence

If the chain is started from a random tree and arbitrarily chosen branch lengths, chances are that the initial likelihood is low. As the chain moves towards the regions in the posterior with high probability mass, the likelihood typically increases very rapidly; in fact, it almost always changes so rapidly that it is necessary to measure it on a log scale (Fig. 7.5). This early phase of the run is known as the burn-in, and the burn-in samples are often discarded because they are so heavily influenced by the starting point.

As the chain approaches its stationary distribution, the likelihood values tend to reach a plateau. This is the first sign that the chain may have converged onto the target distribution. Therefore, the plot of the likelihood values against the generation of the chain, known as the trace plot (Fig. 7.5), is important in monitoring the performance of an MCMC run. However, it is extremely important to confirm convergence using other diagnostic tools because it is not sufficient for the chain to reach the region of high probability in the posterior, it must also cover this region adequately. The speed with which the chain covers the interesting regions of the posterior is known as its mixing behavior. The better the mixing, the faster the chain will generate an adequate sample of the posterior.
Fig. 7.5 The likelihood values typically increase very rapidly during the initial phase of the run because the starting point is far away from the regions in parameter space with high posterior probability. This initial phase of the Markov chain is known as the burn in. The burn-in samples are typically discarded because they are so heavily influenced by the starting point. As the chain converges onto the target distribution, the likelihood values tend to reach a plateau. This has as its hallmark a series of thin horizontal line segments, primarily if the starting point is not chosen carefully. Therefore, the plot of the likelihood values against the generation of the chain, known as the trace plot (Fig. 7.5), is important in monitoring the performance of an MCMC run. However, it is extremely important to confirm convergence using other diagnostic tools because it is not sufficient for the chain to reach the region of high probability in the posterior, it must also cover this region adequately. The speed with which the chain covers the interesting regions of the posterior is known as its mixing behavior. The better the mixing, the faster the chain will generate an adequate sample of the posterior.

samples in this region are discarded!
As the chain approaches its stationary distribution, the likelihood values tend to reach a plateau.

This is the first sign that the chain may have converged onto the target distribution.
However, it is not sufficient for the chain to reach the region of high probability in the posterior; it must also cover this region adequately.

The speed with which the chain covers the interesting regions of the posterior is known as its **mixing behavior**.

The better the mixing, the faster the chain will generate an adequate sample of the posterior.
Fig. 7.6 The time it takes for a Markov chain to obtain an adequate sample of the posterior depends critically on its mixing behavior, which can be controlled to some extent by the proposal tuning parameters. If the proposed values are very close to the current ones, all proposed changes are accepted but it takes a long time for the chain to cover the posterior; mixing is poor. If the proposed values tend to be dramatically different from the current ones, most proposals are rejected and the chain will remain on the same value for a long time, again leading to poor mixing. The best mixing is obtained at intermediate values of the tuning parameters, associated with moderate acceptance rates.

The mixing behavior of a Metropolis sampler can be adjusted using its tuning parameter(s). Assume, for instance, that we are sampling from a normal distribution using a sliding window proposal (Figure 7.6). The sliding window proposal has one tuning parameter, the width of the window. If the width is too small, then the proposed value will be very similar to the current one (Figure 7.6a). The posterior probabilities will also be very similar, so the proposal will tend to be accepted. But each proposal will only move the chain a tiny distance in parameter space, so it will take the chain a long time to cover the entire region of interest; mixing is poor.
In Bayesian MCMC sampling of phylogenetic problems, the tree topology is typically the most difficult parameter to sample from.

Therefore, it makes sense to focus on this parameter when monitoring convergence.
Summarizing the results
The stationary phase of the chain is typically sampled with some thinning, for instance every 50th or 100th generation.

Once an adequate sample is obtained, it is usually trivial to compute an estimate of the marginal posterior distribution for the parameter(s) of interest.
For example, this can take the form of a frequency histogram of the sampled values.

When it is difficult to visualize this distribution or when space does not permit it, various summary statistics are used instead.
The most common approach to summarizing topology posteriors is to give the frequencies of the most common splits, since there are much fewer splits than topologies.
Box 2 | The phylogenetic inference process

The flowchart puts phylogenetic estimation (shown in the green box) into the context of an entire study. After new sequence data are collected, the first step is usually downloading other relevant sequences. Typically, a few outgroup sequences are included in a study to root the tree (that is, to indicate which nodes in the tree are the oldest), provide clues about the early ancestral sequences and improve the estimates of parameters in the model of evolution.

Insertions and deletions obscure which of the sites are homologous. Multiple-sequence alignment is the process of adding gaps to a matrix of data so that the nucleotides (or amino acids) in one column of the matrix are related to each other by descent from a common ancestral residue (a gap in a sequence indicates that the site has been lost in that species, or that a base was inserted at that position in some of the other species). Although models of sequence evolution that incorporate insertions and deletions have been proposed\(^1\)-\(^8\), most phylogenetic methods proceed using an aligned matrix as the input (see REF. 59 for a review of the interplay between alignment and tree inference).

In addition to the data, the scientist must choose a model of sequence evolution (even if this means just choosing a family of models and letting software infer the parameters of these models). Increasing model complexity improves the fit to the data but also increases variance in estimated parameters. Model selection\(^69\)-\(^83\) strategies attempt to find the appropriate level of complexity on the basis of the available data. Model complexity can often lead to computational intractability, so pragmatic concerns sometimes outweigh statistical ones (for example, NJ and parsimony are mainly justifiable by their speed).

As discussed in BOX 3, data and a model can be used to create a sample of trees through either Markov chain Monte Carlo (MCMC) or multiple tree searches on bootstrapped data (the ‘traditional’ approach). This collection of trees is often summarized using consensus-tree techniques, which show the parts of the tree that are found in most, or all, of the trees in a set. Although useful, CONSENSUS METHODS are just one way of summarizing the information in a group of trees. AGREEMENT SUBTREES are more resistant to ‘rogue sequences’ (one or a few sequences that are difficult to place on the tree); the presence of such sequences can make a consensus tree relatively unresolved, even when there is considerable agreement on the relationships between the other sequences.

Sometimes, bootstrap or MCMC sample might show substantial support for multiple trees that are not topologically similar. In such cases, presenting more than one tree (or more than one consensus of trees) might be the only way to appropriately summarize the data.
### Table 1 | Comparison of methods

<table>
<thead>
<tr>
<th>Method</th>
<th>Advantages</th>
<th>Disadvantages</th>
<th>Software</th>
</tr>
</thead>
<tbody>
<tr>
<td>Neighbour joining</td>
<td>Fast</td>
<td>Information is lost in compressing sequences into distances; reliable estimates of pairwise distances can be hard to obtain for divergent sequences</td>
<td>PAUP* MEGA PHYLIP</td>
</tr>
<tr>
<td>Parsimony</td>
<td>Fast enough for the analysis of hundreds of sequences; robust if branches are short (closely related sequences or dense sampling)</td>
<td>Can perform poorly if there is substantial variation in branch lengths</td>
<td>PAUP* NONA MEGA PHYLIP</td>
</tr>
<tr>
<td>Minimum evolution</td>
<td>Uses models to correct for unseen changes</td>
<td>Distance corrections can break down when distances are large</td>
<td>PAUP* MEGA PHYLIP</td>
</tr>
<tr>
<td>Maximum likelihood</td>
<td>The likelihood fully captures what the data tell us about the phylogeny under a given model</td>
<td>Can be prohibitively slow (depending on the thoroughness of the search and access to computational resources)</td>
<td>PAUP* PAML PHYLIP</td>
</tr>
<tr>
<td>Bayesian</td>
<td>Has a strong connection to the maximum likelihood method; might be a faster way to assess support for treeshan maximum likelihood bootstrapping</td>
<td>The prior distributions for parameters must be specified; it can be difficult to determine whether the Markov chain Monte Carlo (MCMC) approximation has run for long enough</td>
<td>MrBayes BAMBE</td>
</tr>
</tbody>
</table>
Acknowledgment

* Material in these slides are based on Chapter 7 in “The Phylogenetic Handbook”, Lemey, Salemi, Vandamme (Eds.)

* Some of the material is based on MCMC notes by Prof. Mark Holder
Questions?