Phylogenetics: Bayesian Phylogenetic Analysis

COMP 571 - Spring 2016
Luay Nakhleh, Rice University
Bayes Rule

\[ P(X = x \mid Y = y) = \frac{P(X = x, Y = y)}{P(Y = y)} = \frac{P(X = x)P(Y = y \mid X = x)}{\sum_{x'} P(X = x')P(Y = y \mid X = x')} \]
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 $p(x=1|y=0)=0.1$. 
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.
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).
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).

![Prior distribution](image)

This is an uninformative prior. The prior distribution assigns equal probability to each of the three possible tree topologies (A, B, and C). After collecting data and using a stochastic evolutionary model and Bayes' theorem, the posterior distribution will reflect the updated probabilities of each tree topology, with the most likely topology receiving the highest probability. If the data are informative, most of the posterior probability will be focused on one tree (or a small subset in a large tree space).
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).

![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).
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

\[ f(X) = \int f(\theta) f(X|\theta) \, d\theta \]

\[ = \sum_{\tau} \int_{v} f(v) f(X|\tau, v) \, dv \]
The marginal probability distribution on topologies
Why are they called marginal probabilities?

![Table](image)

<table>
<thead>
<tr>
<th>Branch length vectors</th>
<th>Topologies</th>
<th>( \tau_A )</th>
<th>( \tau_B )</th>
<th>( \tau_C )</th>
<th>Joint probabilities</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \mathbf{V}^A )</td>
<td>0.10</td>
<td>0.07</td>
<td>0.12</td>
<td></td>
<td>0.29</td>
</tr>
<tr>
<td>( \mathbf{V}^B )</td>
<td>0.05</td>
<td>0.22</td>
<td>0.06</td>
<td></td>
<td>0.33</td>
</tr>
<tr>
<td>( \mathbf{V}^C )</td>
<td>0.05</td>
<td>0.19</td>
<td>0.14</td>
<td></td>
<td>0.38</td>
</tr>
<tr>
<td></td>
<td>0.20</td>
<td>0.48</td>
<td>0.32</td>
<td></td>
<td></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 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 state changes between time step $i$ and time step $i+1$. The transition probabilities are analogous to transition rates if we are working with continuous-time rules for the state of the chain in the next step (step $i+1$ given its current state $x_i$). Note that these probabilistic statements are independent on the state at any point before step $i=0$. In particular, if we are in a particular state at step $i=0$, then the state at step $i+1$ only depends on the state at step $i=0$ and the transition probabilities.

The full probability statements that correspond to the graph are:

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

**What are** $\Pr(x_i = 0 | x_0 = 0)$, $\Pr(x_i = 1 | x_0 = 0)$, $\Pr(x_i = 0 | x_0 = 1)$, $\Pr(x_i = 1 | x_0 = 1)$?
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 $1 = P_j P(x_{i+1} = j | x_i)$ 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:

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

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.

\[
P(x_i = k | x_0 = \ell) = P(x_i = k | x_{i-1} = 0) P(x_{i-1} = 0 | x_0 = \ell) + P(x_i = k | x_{i-1} = 1) P(x_{i-1} = 1 | x_0 = \ell)
\]
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

$$
P(x_{i+1} = j | x_i = k) \text{ for every possible } x_i.$$

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:

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

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 $i$. 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:

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

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.

$$P(x_i = k | x_0 = \ell) = P(x_i = k | x_{i-1} = 0)P(x_{i-1} = 0 | x_0 = \ell) + P(x_i = k | x_{i-1} = 1)P(x_{i-1} = 1 | x_0 = \ell)$$

transition probabilities
Stationary Distribution of a Markov Chain

A Markov chain can be defined by describing the full set of probability statements that define the transition probabilities. The transition probabilities are analogous to transition rates if we are working with continuous-time processes. In general, there are higher-order Markov processes, such as second-order and third-order processes, that depend on the two previous states and the three previous states, respectively. However, in this course, we will only deal 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 transition probabilities. These probability statements are the probability of transitioning from one state to another. For example, the probability of transitioning from state 0 to state 1 is given by the transition probability $P(x_{i+1} = 1 | x_i = 0)$.

Transition probabilities

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

These probabilities are given by the arcs connecting the states with the probabilities next to the line. The states are shown in circles. The transition probabilities are shown as arcs connecting the states with the probabilities next to the line.

Note that the state at step $i + 1$ only depends on the state at step $i$, and the transition probabilities are independent of the history of the chain. So when working with Markov chains, we don't need to concern ourselves with the full history of the chain, merely knowing the current state at the previous step is enough.

$P(x_{i+1} = 1 | x_0, x_1, \ldots, x_i) = P(x_{i+1} = 1 | x_i)$

Clearly if we know $x_i$, then we are dealing with a time-homogeneous Markov chain.

To understand the behavior of the Markov chain, we can consider state changes between steps. Here we are exploiting the fact that the same "rules" (transition probabilities) apply when we consider one-step transitions. But we can also think about the probability that the chain will be in a particular state at step $i$ given its current state (in step $i$). This can be done by making a probabilistic statement $P(x_{i+1} = 1 | x_i = 0)$.

If we consider an even larger number of iterations (e.g. the state at step 100), then the probabilities will stabilize to a steady state distribution. Knowing whether the chain is in a steady state is important in many applications, such as in modeling population dynamics or in predicting future states of a system. Knowing whether the chain will be in a steady state tells you very little about the state of the chain in step 15. Technically, $P(x_{i+1} = 1 | x_i = 0)$ given its current state (in step $i$) is not independent of each other. If you work through the math, the probability of the chain being in state 0 as a function of the step number for the two possible starting states (0 and 1) for the Markov process depicted in Figure (1) stabilizes to a steady state distribution.

Figure 1: A graphical depiction of a two-state Markov process.

Figure 2: The probability of being in state 0 as a function of the step number, for the two possible starting states (0 and 1).

Note that the probability stabilizes to a steady state distribution. Knowing whether the chain is in a steady state is important in many applications, such as in modeling population dynamics or in predicting future states of a system. Knowing whether the chain will be in a steady state tells you very little about the state of the chain in step 15. Technically, $P(x_{i+1} = 1 | x_i = 0)$ given its current state (in step $i$) is not independent of each other. If you work through the math, the probability of the chain being in state 0 as a function of the step number for the two possible starting states (0 and 1) for the Markov process depicted in Figure (1) stabilizes to a steady state distribution.
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 $i$).

Consider the simplest possible Markov chain: one with two states $0$ and $1$. 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:

$$\Pr(x_i = 0 | x_0 = 0)$$

$$\Pr(x_i = 0 | x_0 = 1)$$

The figure to the right shows the states in circles. The transition probabilities are shown $(0, 0.9, 0.1, 0.4, 0.6)$ that operate in discrete time.

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 (0x9 to 792x603).

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:

$$\Pr(x_{15} = 0 | x_0 = 0)$$

$$\Pr(x_{15} = 0 | x_0 = 1)$$

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

But in this course, we'll just be dealing with the simplest Markov chains which only depend on the current state, so clearly these two probabilities are very close to being equal. What this probability statement is saying is that, conditional on $x_i$, $x_{i+1}$ is independent on the state at any point before $x_i$. So when working with Markov processes, the transition probabilities are analogous to transition rates if we are working with continuous-time processes.

Transition probabilities $(0, 0.9, 0.1, 0.4, 0.6)$ can call the state $i$ is positive integer. What this probability statement is saying is that, conditional on $x_i = 0$ or $x_i = 1$, $x_{i+1}$ is independent of each other. If you work through the math the probability of the state at step 15 only depends on the state at step 1.

If the transition probabilities are fixed through the whole duration of the Markov chain, then we are dealing with a time-homogeneous Markov chain.
null
Stationary Distribution of a Markov Chain

A Markov chain can be defined by describing the full set of probability statements that define the transition probabilities. This is the Markov property. More formally we could state it as:

$P(x_{i+1} | x_i) = P(x_{i+1} | x_{i-1}) = \cdots = P(x_{i+k} | x_i) = 0$ for every possible $x_i$, $x_{i+1}, \ldots, x_{i+k}$.

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

$P(x_{i+1} | x_i) = P(x_{i+1} | x_{i-1}) = \cdots = P(x_{i+k} | x_i) = 0$ for every possible $x_i$, $x_{i+1}, \ldots, x_{i+k}$.

But clearly these two probabilities are very close to being equal. 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 transition probabilities. This is the Markov property. More formally we could state it as:

$P(x_{i+1} | x_i) = P(x_{i+1} | x_{i-1}) = \cdots = P(x_{i+k} | x_i) = 0$ for every possible $x_i$, $x_{i+1}, \ldots, x_{i+k}$.

But clearly these two probabilities are very close to being equal.
Stationary Distribution of a Markov Chain

- States: 0 and 1
- Transition Probabilities:
  - From 0 to 0: 0.4
  - From 0 to 1: 0.6
  - From 1 to 0: 0.9
  - From 1 to 1: 0.1

- Stationary Distribution:
  - \( \pi_0 = 0.6 \)
  - \( \pi_1 = 0.4 \)

- Graphical representation of \( P(x_{i+1} = 0 | x_i = 0) \) and \( P(x_{i+1} = 0 | x_i = 1) \)

Where does the 0.6 come from?

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 (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 \( j \) such that:

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

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:

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

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.
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

\[
P_j = \sum_k P(x_{i+1} = j | x_i = k) \quad \text{for every possible } x_i.
\]

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:

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

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 \( i \). 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:

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

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.

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.

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

\[
\pi_0 = \mathbb{P}(x_i = 0) \quad \pi_1 = \mathbb{P}(x_i = 1)
\]
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

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

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:

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

where $k_{i-1}$ 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:

$$P(x_{i+2} = 0 | x_i = 0) = P(x_{i+2} = 0 | x_i = 0, x_{i+1} = 1)P(x_{i+1} = 1 | x_i = 0) + P(x_{i+2} = 0 | x_i = 0, x_{i+1} = 0)P(x_{i+1} = 0 | x_i = 0) = 0.4 + 0.9 = 1$$

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.
Stationary Distribution of a Markov Chain

\[ \frac{\pi_0}{\pi_1} = \frac{\mathbb{P}(x_{i+1} = 0|x_i = 1)}{\mathbb{P}(x_{i+1} = 1|x_i = 0)} \]
Stationary Distribution of a Markov Chain

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

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

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

\[\pi_0 + \pi_1 = 1\]

\[
\frac{\pi_0}{\pi_1} = \frac{0.9}{0.6} = 1.5 \\
\pi_0 = 1.5\pi_1 \\
1.5\pi_1 + \pi_1 = 1.0 \\
\pi_1 = 0.4 \\
\pi_0 = 0.6
\]
Stationary Distribution of a Markov Chain

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:

\[
\begin{align*}
\text{flux out of } j &= \pi_j \mathbb{P}(x_{i+1} \in S \neq j | x_i = j) \\
&= \pi_j [1 - \mathbb{P}(x_{i+1} \in j | x_i = j)] \\
\text{flux into } j &= \sum_{k \in S \neq j} \pi_k \mathbb{P}(x_{i+1} = j | x_i = k) \\
\pi_j [1 - \mathbb{P}(x_{i+1} = j | x_i = j)] &= \sum_{k \in S \neq j} \pi_k \mathbb{P}(x_{i+1} = j | x_i = k) \\
\pi_j &= \pi_j \mathbb{P}(x_{i+1} = j | x_i = j) + \sum_{k \in S \neq j} \pi_k \mathbb{P}(x_{i+1} = j | x_i = k) \\
&= \sum_{k \in S} \pi_k \mathbb{P}(x_{i+1} = j | x_i = k)
\end{align*}
\]
Mixing

While setting the transition probabilities to specific values affects the stationary distribution, the transition probabilities cannot be determined uniquely from the stationary distribution.
\[ P(x_{i+1} = 1 | x_i = 0) = 0.6 \quad P(x_{i+1} = 0 | x_i = 1) = 0.9 \]

\[ P(x_{i+1} = 1 | x_i = 0) = 0.06 \quad P(x_{i+1} = 0 | x_i = 1) = 0.09 \]
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 the posterior probability distribution over the space of parameters. We will try to design chains 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. Starting states (0 and 1) for the Markov process depicted in Figure 1 = 0.09.

Mixing statements as Figure 2 but for a process in which 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 3: The probability of being in state 0 as a function of the step number.

How would such a chain behave? The primary difference is that it would "mix" more slowly. It turns out that the answer is "No." Which can be seen if we examine equation 1.

We just saw how changing transition probabilities will alter the stationary distribution. In the form of a question: if we were given the stationary distribution could we decide what the transition rates must be to achieve that distribution?

There is a one-to-one mapping between transition probabilities and a stationary distribution. In the computer to calculate the probability of being in state 0 for a large number of steps into the future it gets tedious to continue this for a large number of steps into the future. But we can ask a question: if we were given the stationary distribution could we decide what the transition rates are so close that they are indistinguishable.

But clearly these two probabilities are very close to being equal. How would such a chain behave? The primary difference is that it would "mix" more slowly.
Mixing

* 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 $k$:

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

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

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

*desired property*

*P(D) cancels out, so doesn't need to be computed!*

*detailed balance*
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 = \((\text{likelihood ratio}) \cdot (\text{prior ratio}) \cdot (\text{Hastings ratio})\)
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.
Fig. 7.4 The Markov chain Monte Carlo (MCMC) procedure is used to generate a valid sample from the posterior. One first sets up a Markov chain that has the posterior as its stationary distribution. The chain is then started at a random point and run until it converges onto this distribution. In each step (generation) of the chain, a small change is made to the current values of the model parameters (step 2). The ratio $r$ of the posterior probability of the new and current states is then calculated. If $r > 1$, we are moving uphill and the move is always accepted (3a). If $r < 1$, we are moving downhill and accept the new state with probability $r$ (3b).

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 **beta proposal**:

1. Assume the current values are \((x_1, x_2)\);
2. Multiply them with a value \(\alpha\);
3. Pick new values from \(\text{Beta}(\alpha x_1, \alpha x_2)\);
4. Pick new values from \(\text{Beta}(\alpha x_1, \alpha x_2)\)
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.
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 something to do with the fact that the Markov chain has something to do with the data.

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 so far been a phase dwilts with something, primarily to savedisk space.

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.

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 (Fig. 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 (Fig. 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 (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 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, N 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, the 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.

Summary
## Summary

**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 treesthan 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>

*Source: Nat Rev Genet, 4:275, 2003*
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?