Bayesian Phylogenetics

Bayes Theorem

- \( \Pr(\text{Tree}|\text{Data}) = \frac{\Pr(\text{Data}|\text{Tree}) \times \Pr(\text{Tree})}{\Pr(\text{Data})} \)

- \( \Pr(\text{Tree}) = \text{Prior probability of the tree} \)

- \( \Pr(\text{Data}) = \text{Prior probability of the data} \)
  - \( \Pr(\text{Data}|\text{Tree}) \) over all trees, weighted by their prior probabilities

- \( \Pr(\text{Data}|\text{Tree}) = \text{Likelihood of the data given the tree} \)

- \( \Pr(\text{Tree}|\text{Data}) = \text{Posterior probability of the tree} \)

Assuming the prior probability of two trees (i.e., tree topologies) equal, the ratio of their posterior probabilities equals the ratio of their likelihood scores. True or false?

\[
\begin{align*}
\Pr(\text{Tree}_1|\text{Data}) &= \frac{\Pr(\text{Data}|\text{Tree}_1) \times \Pr(\text{Tree}_1)}{\Pr(\text{Data})} \\
\Pr(\text{Tree}_2|\text{Data}) &= \frac{\Pr(\text{Data}|\text{Tree}_2) \times \Pr(\text{Tree}_2)}{\Pr(\text{Data})} \\
\frac{\Pr(\text{Tree}_1|\text{Data})}{\Pr(\text{Tree}_2|\text{Data})} &= \left( \frac{\Pr(\text{Data}|\text{Tree}_1) \times \Pr(\text{Tree}_1)}{\Pr(\text{Data})} \right) \left( \frac{\Pr(\text{Data})}{\Pr(\text{Data}|\text{Tree}_2) \times \Pr(\text{Tree}_2)} \right)
\end{align*}
\]
Assuming the prior probability of two trees (i.e., tree topologies) equal, the ratio of their posterior probabilities equals the ratio of their likelihood scores. True or false?

\[
\frac{\Pr(T_{\text{Tree1}}|\text{Data})}{\Pr(T_{\text{Tree2}}|\text{Data})} = \frac{\Pr(\text{Data}|T_{\text{Tree1}}) \times \Pr(T_{\text{Tree1}})}{\Pr(\text{Data}|T_{\text{Tree2}}) \times \Pr(T_{\text{Tree2}})}
\]

If tree topology is the parameter of interest, what are some “nuisance parameters” that need to be accommodated in a Bayesian or Maximum Likelihood analysis?

- Branch lengths
- Substitution rates
- Base frequencies
- Rate heterogeneity parameters

How do Bayesian and Likelihood approaches differ in their treatment of nuisance parameters?

- **Likelihood**
  - Find the value of each parameter that maximizes the likelihood of the data

- **Bayesian**
  - Integrate over all possible values of each parameter (weighted by the prior probability distribution)

How does MCMC get around the problem of not being able to calculate \(\Pr(\text{Data})\)?

- The probability of staying/leaving a tree is determined by its posterior probability relative to other nearby trees
- The overall time spent on a tree will converge to its absolute posterior probability

Like walking-over tree space at a rate governed by altitude..

What happens at each “step” of a Bayesian, phylogenetic MCMC analysis?

- A new parameter (topology, branch lengths, substitution parameters etc.) is proposed
- Whether the new parameter is accepted is governed by the metropolis-hasting equation

\[
\min \left\{ 1, \frac{\Pr(\text{Data}|x = i)}{\Pr(\text{Data}|x = \cdot)} \times \frac{\Pr(x = i)}{\Pr(x = \cdot)} \times \frac{\Pr(\text{proposing } x = i | x = \cdot)}{\Pr(\text{proposing } x = \cdot | x = i)} \right\}
\]

**Posterior ratio** **Prior ratio** **Proposal ratio**
Why is the output of MCMC called a “posterior distribution?” What does it contain?

- It contains a list of parameters in effect at a sampled set of generations (after burnin)
- The frequency of a parameter in this sample should be proportional to its posterior probability

How can the posterior distribution be queried to find the posterior probability of a clade?

- Just see the proportion of trees in the distribution that have the clade

How can the posterior distribution be queried to evaluate other parameters, for example the transition:transversions ratio?

- Generate a histogram (or fit to a probability density function)
- Establish a credibility interval – the range that encompasses, say, 95% of the distribution

A dice example

- A manufacturer makes regular dice and trick dice (with 2 sixes) – in equal numbers
- You are given a die from this manufacturer and are not allowed to look at all the sides – you can only look at the side that is up after a roll
- You roll this die 10 times and get these numbers

• What is the probability that the die is one with two sixes?

Bayesian approach

\[
Pr(H|D) = \frac{Pr(D|H) \times Pr(H)}{Pr(D)}
\]

<table>
<thead>
<tr>
<th>1-six</th>
<th>2-six</th>
<th>Sum</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pr(H)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Pr(D</td>
<td>H)</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>1-six</th>
<th>2-six</th>
<th>Sum</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pr(H)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Pr(D</td>
<td>H)</td>
<td></td>
</tr>
</tbody>
</table>

Bayesian approach

\[
Pr(H|D) = \frac{Pr(D|H) \times Pr(H)}{Pr(D)}
\]
Bayesian approach

$$\Pr(H|D) = \frac{\Pr(D|H) \times \Pr(H)}{\Pr(D)}$$

<table>
<thead>
<tr>
<th></th>
<th>1-six</th>
<th>2-six</th>
<th>Sum</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pr(H)</td>
<td>0.5</td>
<td>0.5</td>
<td>1.0</td>
</tr>
<tr>
<td>Pr(D</td>
<td>H)</td>
<td>7.44E-05</td>
<td>0.00123</td>
</tr>
</tbody>
</table>

MCMC approach

$$\min \left\{ 1, \frac{\Pr(\text{Data}|x = i) \times 1 \times 1 \times 1}{\Pr(\text{Data}|x = i)} \right\}$$

Bayesian approach

$$\Pr(H|D) = \frac{\Pr(D|H) \times \Pr(H)}{\Pr(D)}$$

<table>
<thead>
<tr>
<th></th>
<th>1-six</th>
<th>2-six</th>
<th>Sum</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pr(H)</td>
<td>0.5</td>
<td>0.5</td>
<td>1.0</td>
</tr>
<tr>
<td>Pr(D</td>
<td>H)</td>
<td>(1/6)^5*(5/6)^3</td>
<td>(2/6)^5*(4/6)^3</td>
</tr>
</tbody>
</table>

MCMC approach

$$\min \left\{ 1, \frac{\Pr(\text{Data}|\text{one six})}{\Pr(\text{Data}|\text{two sixes})} \right\} = 1$$
If you run this long enough, you will spend only 4.7% of the time on “1 six.” This is the PP of that hypothesis.