Bayesian Phylogenetics

Paul O. Lewis
Department of Ecology & Evolutionary Biology
University of Connecticut

Smithsonian Molecular Evolution Workshop
3 June 2009
An Introduction to Bayesian Phylogenetics

• Bayesian inference in general
• Markov chain Monte Carlo (MCMC)
• Bayesian phylogenetics
• Prior distributions
• Bayesian model selection
I. Bayesian inference in general
Joint probabilities

\[ B = \text{Black} \quad S = \text{Solid} \]
\[ W = \text{White} \quad D = \text{Dotted} \]

\[
\begin{align*}
\Pr(B) &= 0.6 \\
\Pr(W) &= 0.4 \\
\Pr(S) &= 0.5 \\
\Pr(D) &= 0.5 \\
\Pr(\bullet) &= \Pr(B, D) = 0.2 \\
\Pr(\circ) &= \Pr(B, S) = 0.4 \\
\Pr(\circledcirc) &= \Pr(W, D) = 0.3 \\
\Pr(\circledcirc) &= \Pr(W, S) = 0.1 
\end{align*}
\]
Conditional probabilities

\[ \Pr(B|D) = \frac{2}{5} = 0.4 \]

Hide all solid marbles (leaving 5 with dot)

Of those left, 2 are black
Bayes’ rule

\[
Pr(B, D) = Pr(D) \cdot Pr(B|D) = Pr(B) \cdot Pr(D|B)
\]

\[
\frac{1}{2} \times \frac{2}{5} = \frac{3}{5} \times \frac{1}{3}
\]

\[
Pr(B|D) = \frac{Pr(B) \cdot Pr(D|B)}{Pr(D)}
\]

\[
= \frac{\frac{3}{5} \times \frac{1}{3}}{\frac{1}{2}} = \frac{2}{5}
\]
Probability of "Dotted"

\[ \Pr(D) = \Pr(D,W) + \Pr(D,B) \]
Bayes' rule (cont.)

\[ \Pr(B|D) = \frac{\Pr(B) \Pr(D|B)}{\Pr(D)} = \frac{\Pr(B) \Pr(D|B)}{\Pr(D, B) + \Pr(D, W)} \]

\( \Pr(D) \) is the **marginal probability** of being dotted
To compute it, we **marginalize over colors**
Joint probabilities

<table>
<thead>
<tr>
<th></th>
<th>B</th>
<th>W</th>
</tr>
</thead>
<tbody>
<tr>
<td>D</td>
<td>Pr(D,B)</td>
<td>Pr(D,W)</td>
</tr>
<tr>
<td>S</td>
<td>Pr(S,B)</td>
<td>Pr(S,W)</td>
</tr>
</tbody>
</table>
Marginalizing over colors

Marginal probability of a dotted marble is the sum of all joint probabilities involving dotted marbles.
## Marginal probabilities

<table>
<thead>
<tr>
<th></th>
<th>B</th>
<th>W</th>
</tr>
</thead>
<tbody>
<tr>
<td>D</td>
<td></td>
<td></td>
</tr>
<tr>
<td>S</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

- **Marginal probability of being dotted**: $\Pr(D,B) + \Pr(D,W)$
- **Marginal probability of being solid**: $\Pr(S,B) + \Pr(S,W)$
Marginalizing over "dottedness"

\[ \Pr(D, B) \]
\[ \Pr(S, B) \]
\[ \Pr(D, W) \]
\[ \Pr(S, W) \]

Marginal probability of being a white marble
Bayes' rule (cont.)

$$\Pr(B|D) = \frac{\Pr(B) \Pr(D|B)}{\Pr(D, B) + \Pr(D, W)}$$

$$= \frac{\Pr(B) \Pr(D|B)}{\Pr(B) \Pr(D|B) + \Pr(W) \Pr(D|W)}$$

$$= \frac{\Pr(B) \Pr(D|B)}{\sum_{\theta \in \{B, W\}} \Pr(\theta) \Pr(D|\theta)}$$
Bayes' rule in Statistics

\[
\Pr(\theta \mid D) = \frac{\Pr(D \mid \theta) \Pr(\theta)}{\sum_{\theta} \Pr(D \mid \theta) \Pr(\theta)}
\]

\(D\) refers to the "observables" (i.e. the Data)

\(\theta\) refers to one or more "unobservables"

(i.e. parameters of a model, or the model itself):

– tree model (i.e. tree topology)
– substitution model (e.g. JC, F84, GTR, etc.)
– parameter of a substitution model (e.g. a branch length, a base frequency, transition/transversion rate ratio, etc.)
– hypothesis (i.e. a special case of a model)
– a latent variable (e.g. ancestral state)
Bayes’ rule in statistics

\[ \text{Pr}(\theta|D) = \frac{\text{Pr}(D|\theta) \cdot \text{Pr}(\theta)}{\sum_{\theta} \text{Pr}(D|\theta) \cdot \text{Pr}(\theta)} } \]

- **Likelihood** of hypothesis \( \theta \)
- **Prior probability** of hypothesis \( \theta \)
- **Posterior probability** of hypothesis \( \theta \)
- **Marginal probability of the data** (marginalizing over hypotheses)
**Simple (albeit silly) paternity example**

$\theta_1$ and $\theta_2$ are assumed to be the only possible fathers, child has genotype $Aa$, mother has genotype $aa$, so child must have received allele $A$ from the true father. Note: the data in this case is the child's genotype ($Aa$)

<table>
<thead>
<tr>
<th>Possibilities</th>
<th>$\theta_1$</th>
<th>$\theta_2$</th>
<th>Row sum</th>
</tr>
</thead>
<tbody>
<tr>
<td>Genotypes</td>
<td>AA</td>
<td>Aa</td>
<td>---</td>
</tr>
<tr>
<td>Prior</td>
<td>1/2</td>
<td>1/2</td>
<td>1</td>
</tr>
<tr>
<td>Likelihood</td>
<td>1</td>
<td>1/2</td>
<td>---</td>
</tr>
<tr>
<td>Prior X Likelihood</td>
<td>1/2</td>
<td>1/4</td>
<td>3/4</td>
</tr>
<tr>
<td>Posterior</td>
<td>2/3</td>
<td>1/3</td>
<td>1</td>
</tr>
</tbody>
</table>

© 2009 Paul O. Lewis
Discrete vs. Continuous

• So far, we've been dealing with discrete hypotheses (e.g. either this father or that father)

• In phylogenetics, substitution models represent an infinite number of hypotheses (each combination of parameter values is in some sense a separate hypothesis)

• How do we use Bayes' rule when our hypotheses form a continuum?
If you had to guess...

Not knowing anything about my archery abilities, draw a curve representing your view of the chances of my arrow landing a distance $d$ from the center of the target (if it helps, I'm standing 50 meters away from the target).
Case 1: assume I have talent

An informative prior (low dispersion = low variance) that says most of my arrows will fall within 10 cm of the center (thanks for your confidence!)
Case 2: assume I have a talent for missing the target!

Also an *informative* prior, but one that says most of my arrows will fall within a narrow range just outside the entire target!
Case 3: assume I have no talent

This is a vague prior: its high variance reflects nearly total ignorance of my abilities, saying that my arrows could land nearly anywhere!
Notice that I haven't provided a scale for the vertical axis.

What exactly does the height of this curve mean?

Question: does the height of the dotted line represent the *probability* that my arrow lands 60 cm from the center of the target?
Probabilities apply to intervals

Probabilities are attached to intervals (i.e. ranges of values), not individual values.

The probability of any given point (e.g. \(d = 60.0\)) is zero!

However, we can ask about the probability that \(d\) falls in a particular range, e.g. \(50.0 < d < 65.0\).
The type of curve we have been discussing is known as a **probability density function**.

The height of the curve at any given value represents the *probability density* at that value, *not* the *probability* of the value.

The value 2.0 has low probability density compared to the value 1.0.
Probabilities vs. probability densities

Probability *density* function
Note: the height of this curve does *not* represent a probability (if it did, it would not exceed 1.0)
# Densities of various substances

<table>
<thead>
<tr>
<th>Substance</th>
<th>Density (g/cm³)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cork</td>
<td>0.24</td>
</tr>
<tr>
<td>Aluminum</td>
<td>2.70</td>
</tr>
<tr>
<td>Gold</td>
<td>19.30</td>
</tr>
</tbody>
</table>

*Density does not equal mass*

\[
\text{mass} = \text{density} \times \text{volume}
\]

Note: *volume* is appropriate for 3-dimensional objects or materials.
For 2-dimensions, *area* take the place of volume.
For 1-dimension, *linear distance* replaces volume.
Can say: values in the vicinity of 1.24 are twice as probable as values in the vicinity of 1.75

Cannot say: probability of the exact value 1.24 is 0.4
Integration of densities

The density curve is scaled so that the value of this integral (i.e. the total area) equals 1.0
Integration of a probability density yields a probability

\[ \int_0^2 f(\theta) \, d\theta \]

Area under the density curve from 0 to 2 is the probability that \( \theta \) is less than 2.
Archery Priors Revisited

These density curves are all variations of a **gamma probability distribution**. We could have used a gamma distribution to specify each of the prior probability distributions for the archery example. Note that **higher variance means less informative**.
Coin-flipping

\[ y = \text{observed number of heads} \]
\[ n = \text{number of flips (sample size)} \]
\[ p = \text{(unobserved) proportion of heads} \]

\[
Pr(y \mid p) = \frac{n!}{y!(n-y)!} p^y (1 - p)^{n-y} \quad \text{The same formula is the:}
\]

\[
L(p \mid y) = \frac{n!}{y!(n-y)!} p^y (1 - p)^{n-y} \quad \text{probability of } y \text{ (if } p \text{ is fixed)}
\]

\[
\text{likelihood of } p \text{ (if } y \text{ is fixed)}
\]
### Likelihood vs. Probability

<table>
<thead>
<tr>
<th>Outcome</th>
<th>Fair coin model</th>
<th>Two-heads model</th>
</tr>
</thead>
<tbody>
<tr>
<td>H</td>
<td>0.5</td>
<td>1.0</td>
</tr>
<tr>
<td>T</td>
<td>0.5</td>
<td>0.0</td>
</tr>
</tbody>
</table>

- **Probabilities** are functions of the data (the model is fixed)
- **Likelihoods** are functions of models (the data is fixed)

- Symbol $\Pr(D|M)$ often used to represent the likelihood (I will be following convention in this regard)
- Say "likelihood of the model", not "likelihood of the data"
The posterior is more informative than the prior (data contains information)

posterior density

uniform prior density

= posterior probability (mass)
Bayes’ rule: continuous case

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

- Likelihood
- Prior probability density
- Posterior probability density
- Marginal probability of the data
Beta prior gives more flexibility

Beta(2,2) prior density

posterior density

Posterior probability of $p$ between 0.45 and 0.55 is 0.223

Note:
- prior is vague but not flat
- posterior has smaller variance (i.e. it is more informative) than the prior
Bayesian Coin Flipper

Windows program download from:
http://www.eeb.uconn.edu/people/plewis/software.php
Usually there are many parameters...

A 2-parameter example

\[ f(\theta, \phi | D) = \frac{\int f(D | \theta, \phi) f(\theta) f(\phi) d\theta d\phi}{\int \int f(D | \theta) f(\theta) f(\phi) d\theta d\phi} \]

Prior probability density

Likelihood

Posterior probability density

Marginal probability of data

An analysis of 100 sequences under the simplest model (JC69) requires 197 branch length parameters. The denominator is a 197-fold integral in this case! Now consider summing over all possible tree topologies! It would thus be nice to avoid having to calculate the marginal probability of the data...
II. Markov chain Monte Carlo (MCMC)
Markov chain Monte Carlo (MCMC)

For more complex problems, we might settle for a good approximation to the posterior distribution.
MCMC robot’s rules

- Uphill steps are always accepted.
- Slightly downhill steps are usually accepted.
- Drastic “off the cliff” downhill steps are almost never accepted.

With these rules, it is easy to see why the robot tends to stay near the tops of hills.
(Actual) MCMC robot rules

Uphill steps are always accepted because \( R > 1 \)

Slightly downhill steps are usually accepted because \( R \) is near 1

Drastic “off the cliff” downhill steps are almost never accepted because \( R \) is near 0

The robot takes a step if it draws a Uniform(0,1) random deviate that is less than or equal to \( R \)
Cancellation of marginal likelihood

When calculating the ratio $R$ of posterior densities, the marginal probability of the data cancels.

$f(D)$ is a constant for any particular data set. It is the probability of the data taking into account all possible values of $\theta$, hence it does not depend on any particular value of $\theta$.

\[
\frac{f(\theta^* | D)}{f(\theta | D)} = \frac{\frac{f(D | \theta^*) f(\theta^*)}{f(D)}}{\frac{f(D | \theta) f(\theta)}{f(D)}} = \frac{f(D | \theta^*) f(\theta^*)}{f(D | \theta) f(\theta)}
\]

Posterior odds

Likelihood ratio

Prior odds
Target vs. proposal distributions

• The **target distribution** is the posterior distribution of interest

• The **proposal distribution** is used to decide which point to try next
  – you have much flexibility here, and the choice affects only the **efficiency** of the MCMC algorithm
  – MCMC using a **symmetric** proposal distribution is the Metropolis algorithm (Metropolis et al. 1953)
  – Use of an **asymmetric** proposal distribution requires a modification proposed by Hastings (1970), and is known as the Metropolis-Hastings algorithm

Target vs. Proposal Distributions

Pretend this proposal distribution allows **good mixing**. What happens if we change it?
Trace plots

“White noise” appearance is a sign of good mixing

I used the program Tracer to create this plot:
http://tree.bio.ed.ac.uk/software/tracer/

AWTY (Are We There Yet?) is useful for investigating convergence:
http://king2.scs.fsu.edu/CEBProjects/awty/awty_start.php
Target vs. Proposal Distributions

Proposal distributions with smaller variance...

Advantage: robot seldom refuses to take proposed steps

Disadvantage: robot takes smaller steps, more time required to explore the same area
If step size is too small, large-scale trends will be apparent.
Target vs. Proposal Distributions

Proposal distributions with larger variance...

Disadvantage: robot often proposes a step that would take it off a cliff, and refuses to move

Advantage: robot can potentially cover a lot of ground quickly
Chain is spending long periods of time “stuck” in one place.

“Stuck” robot is indicative of step sizes that are too large (most proposed steps would take the robot “off the cliff”).
The Hastings ratio

If robot has a greater tendency to propose steps to the right as opposed to the left when choosing its next step, then the acceptance ratio must counteract this tendency.

Suppose the probability of proposing a spot to the right is 2/3 (making the probability of choosing left 1/3).

In this case, the Hastings ratio decreases the chance of accepting moves to the right by half, and increases the chance of accepting moves to the left (by a factor of 2), thus exactly compensating for the asymmetry in the proposal distribution.

Hastings Ratio

\[ R = \left[ \frac{f(D|\theta^*) f(\theta^*)}{f(D|\theta) f(\theta)} \right] \left[ \frac{q(\theta|\theta^*)}{q(\theta^*|\theta)} \right] \]

Acceptance ratio  Posterior ratio  Hastings ratio

Note that if \( q(\theta|\theta^*) = q(\theta^*|\theta) \), the Hastings ratio is 1
MCRobot

Windows program download from:
http://www.eeb.uconn.edu/people/plewis/software.php
What MCRobot can teach us about Markov chain Monte Carlo

Proposal scheme:
• random direction
• gamma-distributed step length
• reflection at edges

Posterior distribution:
• equal mixture of 3 bivariate normal “hills”
• inner contours: 50%
• outer contours: 95%

Here you are looking down from above at one of the three bivariate normal hills

© 2009 Paul O. Lewis
Random walk

In this case, the robot is accepting every step.

5000 steps are shown. The robot left a dot each place it stepped.

Density of points does not reflect probability density (that is, a pure random walk does not help us)
First 100 steps

Robot is now following the rules and thus quickly finds one of the three hills

Note that first few steps are not at all representative of the distribution (this is the “burn-in” period)
Target distribution approximation

How good is the MCMC approximation?

- 51.2% of points are inside inner contours (cf. 50% actual)
- 93.6% of points are inside outer contours (cf. 95% actual)

Approximation gets better the longer the chain is allowed to run.
Just how long is a long run?

What would you conclude about the target distribution had you stopped the robot at this point?

The way to avoid this mistake is to perform several runs, each one beginning from a different randomly-chosen starting point.

Results different among runs? Probably none of them were long enough!
Tradeoff

- Taking **big steps** helps in jumping from one “island” in the posterior density to another
- Taking **small steps** often results in better mixing
- How can we overcome this tradeoff? **MCMC**
Metropolis-coupled Markov chain Monte Carlo (MCMC)

- MCMC involves running **several chains simultaneously**
- The **cold chain** is the one that counts, the rest are **heated chains**
- Chain is heated by raising densities to a power less than 1.0 (values closer to 0.0 are warmer)

Heated chains act as scouts for the cold chain

© 2009 Paul O. Lewis
Cold and hot chains swapped
Cold vs. heated landscapes

Cold landscape (this is what is being explored by the cold chain): note peaks separated by deep valleys

Heated landscape (this is what is being explored by one of the heated chains): note the shallow valleys (easy to cross)
III. Bayesian phylogenetics
So, what’s all this got to do with phylogenetics?

Imagine pulling out trees at random from a barrel. In the barrel, some trees are represented numerous times, while other possible trees are not present. Count 1 each time you see the split separating just A and C from the other taxa, and count 0 otherwise. Dividing by the total trees sampled approximates the true proportion of that split in the barrel.
Moving through treespace

The Larget-Simon* move

Step 1: select 3 contiguous branch segments (bolded)

Step 2: shrink or expand selected segment by a random amount

\[ m^* = m e^{\lambda(u - 1/2)} \]

Step 3: select one of 2 groups attached to selected segment at random and prune (group X selected here)

Step 4: reattach pruned group to selected segment at a random point (this will change topology of tree if reattachment occurs in this region)


This shows the tree after the proposed move has been accepted. The selected segment has been shortened, and group X ended up on a different segment, thus changing the topology.
Moving through parameter space

Using $\kappa$ (ratio of the transition rate to the transversion rate) as an example of a model parameter.

Proposal distribution is the uniform distribution on the interval $(\kappa-d, \kappa+d)$

The “step size” of the MCMC robot is defined by $d$: a larger $d$ means that the robot will attempt to make larger jumps on average.
Putting it all together

- **Start with** random tree and arbitrary initial values for branch lengths and model parameters
- **Each generation** consists of one of these (chosen at random):
  - Propose a **new tree** (e.g. Larget-Simon move) and either accept or reject the move
  - Propose (and either accept or reject) a **new model parameter value**
- Every $k$ generations, save tree topology, branch lengths and all model parameters (i.e. sample the chain)
- After $n$ generations, **summarize sample** using histograms, means, credible intervals, etc.
Marginal Posterior Distribution of $\kappa$

Histogram created from a sample of 1000 kappa values.

lower = 2.907  upper = 3.604

95% credible interval

mean = 3.234

IV. Prior distributions
Common Priors

- **Discrete uniform** for topologies
  - exceptions becoming more common
- **Beta** for proportions
- **Gamma** for branch lengths and other parameters with support $[0,\infty)$
  - Exponential is common special case
- **Dirichlet** for state frequencies and GTR
Discrete Uniform distribution for topologies

\[
\begin{array}{cccc}
A & E & C & B \\
D & B & & \\
\hline \\
A & B & E & D \\
C & & & B \\
\hline \\
A & C & B & D \\
E & & & A \\
\hline \\
B & C & A & E \\
D & & & B \\
\hline \\
B & D & C & E \\
A & & & D \\
\hline \\
B & E & D & C \\
A & & & B \\
\hline \\
\end{array}
\]
The rate of speciation under the Yule model ($\lambda$) is constant and applies equally and independently to each lineage. Thus, speciation events get closer together in time as the tree grows because more lineages are available to speciate.
**Gamma**($a,b$) distribution for **branch lengths** and parameters with support $[0,\infty)$

\[ f(x) = \frac{x^{a-1} e^{-x/b}}{b^a \Gamma(a)} \]

- $a$ is the **shape** parameter
- $b$ is the **scale** parameter
- mean = $ab$
- variance = $ab^2$
- $a = 1$ is a special case: Exponential($1/b$) distribution
- $\Gamma(x) = (x-1)!$ (if $x$ is a positive whole number)

Note: be aware that in many papers the Gamma distribution is defined such that the second (scale) parameter is the inverse of the value $b$ used in this slide! In this case, the mean and variance would be $a/b$ and $a/b^2$, respectively.
Gamma($a,b$) Gallery

- **Gamma(0.1, 10)**: shoots off to infinity
- **Exponential(1) = Gamma(1, 1)**: hits y-axis at $b$ if $a = 1$
- **Gamma(400, 0.01)**: peak > 0 if $a > 1$
Beta$(a,b)$ distribution for proportions

$$f(p) = \frac{p^{a-1}(1-p)^{b-1}}{\frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)}}$$

this ratio of gamma functions is the normalizing constant

- $a$ and $b$ are the two shape parameters
- mean = $a/(a + b)$
- variance = $ab/[(a + b)^2 (a + b + 1)]$
- $a = b = 1$ is a special case: Uniform$(0,1)$ distribution
Beta\((a,b)\) Gallery

Beta\((0.8,2)\)
- leans left if \(a < b\)
- mean = \(a/(a+b)\)

Beta\((10,10)\)
- symmetric if \(a = b\)
- mean = \(a/(a+b) = 0.5\)

Beta\((1.2,2)\)
- flat if \(a = b = 1\)

Beta\((1,1)\)
**Dirichlet**\((a,b,c,d)\) for base frequencies

**Dirichlet**\((a,b,c,d,e,f)\) for GTR relative rates

\[ f(p) = \left( \frac{\Gamma(a + b)}{\Gamma(a)\Gamma(b)} \right) p^{a-1}(1 - p)^{b-1} \]

**Dirichlet**\((a,b,c)\) density function:

\[ f(p, q) = \left( \frac{\Gamma(a + b + c)}{\Gamma(a)\Gamma(b)\Gamma(c)} \right) p^{a-1}q^{b-1}(1 - p - q)^{c-1} \]

**Dirichlet**\((a,b,c,d)\) density function:

\[ f(p, q, r) = \left( \frac{\Gamma(a + b + c + d)}{\Gamma(a)\Gamma(b)\Gamma(c)\Gamma(d)} \right) p^{a-1}q^{b-1}r^{c-1}(1 - p - q - r)^{d-1} \]
Dirichlet($a,b,c,d$) distribution

Flat prior:
$a = b = c = d = 1$

Informative prior:
$a = b = c = d = 300$

(stereo pairs)
How much information is present in data about nucleotide frequencies?

Data:
- 26S rDNA gene
- 34 species of green algae
- 1688 aligned nucleotide sites

Analysis:
- 2,000,000 generations
- last 1000 samples (of 2000 total)
- HKY+gamma model
Dirichlet($a,b,c,d,e,f$) distribution

- $a\rightarrow r_{AC}$, $b\rightarrow r_{AG}$, $c\rightarrow r_{AT}$, $d\rightarrow r_{CG}$, $e\rightarrow r_{CT}$, $f\rightarrow r_{GT}$
- flat when $a = b = c = d = e = f = 1$
- an informative prior with all rates nearly equal: $a = b = c = d = e = f = 300$
- a vague prior where rate of transitions slightly higher than the rate of transversions: $a = c = d = f = 1$ and $b = e = 2$
- mean for $r_{AC}$ is $a/s$ where $s = a + b + c + d + e + f$
- variance for $r_{AC}$ is $a(s - a) / [s^2(s + 1)]$
- Beta($a,b$) equals Dirichlet($a,b$)
Prior Miscellany
This Gamma(4,1) prior ties down its parameter at the mode, which is at 3, and discourages it from venturing too far in either direction. For example, a parameter value of 10 would be stretching the rubber band fairly tightly.

The mode of a Gamma($a, b$) distribution is $(a-1)b$ (assuming $a > 1$).
This Gamma prior also has a mode at 3, but has a variance 40 times smaller. Decreasing the variance is tantamount to increasing the strength of the metaphorical rubber band.

Now the parameter must struggle hard just to reach a value as small as 2 or as large as 4.

This gamma distribution has shape 91.989 and scale 0.032971
Too many parameters, too little information

7 coins flipped once

H H T T T T H

3/7 = 0.43
(maximum likelihood estimate)

1.0 1.0 0.0 0.0 0.0 0.0 1.0
(maximum likelihood estimates)

1 parameter model

7 parameter model

Maximum likelihood estimates go to extremes if there is too little information in the data. It is clear that data from only one flip does not provide enough information to estimate the proportion of heads accurately.
Priors supplement information in data

H H T T T T H

5/11 = 0.455
(posterior mean)

0.6 0.6 0.4 0.4 0.4 0.4 0.6
(posterior means)

7 coins flipped once

1 parameter model

7 parameter model

Under **Bayesian inference**, priors **add information** and can keep complex models in check. Here I applied a Beta(2,2) prior to each parameter. This did not change the 1 parameter results by much, but the 7-parameter results are now much more reasonable.

**Question:** Where did I pound in the stake used to tie down the parameter values?
In this case, the prior has the stronger rubber band (the posterior mean is 0.6, which is closer to 0.5 than 1.0).
The choice of prior distributions can potentially turn a good model bad!

LRT, AIC and BIC all say this is a great model because it is able to attain such a high maximum likelihood score. But the prior never allows the parameter out of this box, so in actuality the model performs very poorly.
Example: Internal Branch Length Priors

Internal branch length prior mean 0.1

This is a reasonably vague internal branch length prior
Internal branch length prior mean 0.01

Not much effect yet...
Internal branch length prior mean 0.001

Notice how the internal branch lengths are shrinking...

(Trees in this series are drawn to same scale)
Internal branch length prior mean 0.0001

Model compensating for small internal branch lengths by increasing the external branch lengths...
Internal branch length prior mean 0.00001

Internal branch length prior now so informative that it is beginning to noticeably override the likelihood...
The internal branch length prior is calling the shots now.
Running on empty

**Solid line:** prior density estimated from MrBayes output

**Dotted line:** exponential(10) density for comparison

You can use the program Tracer to show the estimated density:
http://tree.bio.ed.ac.uk/software/tracer/
In a non-hierarchical model, all parameters are present in the likelihood function

Prior: Exponential, mean=0.1

\[ L_k = \frac{1}{4} \left[ \frac{1}{4} + \frac{3}{4} e^{-4\nu_1/3} \right] \left[ \frac{1}{4} + \frac{3}{4} e^{-4\nu_2/3} \right] \left[ \frac{1}{4} - \frac{1}{4} e^{-4\nu_3/3} \right] \left[ \frac{1}{4} - \frac{1}{4} e^{-4\nu_4/3} \right] \left[ \frac{1}{4} + \frac{3}{4} e^{-4\nu_5/3} \right] \]
Hierarchical models add hyperparameters not present in the likelihood function.

\[ L_k = \frac{1}{4} \left[ \frac{1}{4} + \frac{3}{4} e^{-4v_1/3} \right] \left[ \frac{1}{4} + \frac{3}{4} e^{-4v_2/3} \right] \left[ \frac{1}{4} - \frac{1}{4} e^{-4v_3/3} \right] \left[ \frac{1}{4} - \frac{1}{4} e^{-4v_4/3} \right] \left[ \frac{1}{4} + \frac{3}{4} e^{-4v_5/3} \right] \]

During an MCMC analysis, \( \mu \) will hover around a reasonable value, sparing you from having to decide what value is appropriate. You still have to specify a hyperprior, however.
Empirical Bayes

Empirical Bayes uses the data to determine some aspects of the prior, such as the prior mean. This uses the data twice, which is not acceptable to Bayesian purists.

An empirical Bayesian would use the maximum likelihood estimate (MLE) of the length of an average branch here.

Prior: Exponential, mean=MLE

\[
L_k = \frac{1}{4} \left[ \frac{1}{4} + \frac{3}{4} e^{-4V_1/3} \right] \left[ \frac{1}{4} + \frac{3}{4} e^{-4V_2/3} \right] \left[ \frac{1}{4} - \frac{1}{4} e^{-4V_3/3} \right] \left[ \frac{1}{4} - \frac{1}{4} e^{-4V_4/3} \right] \left[ \frac{1}{4} + \frac{3}{4} e^{-4V_5/3} \right]
\]
V. Bayesian model selection
Marginal probabilities of models

\[
\Pr(D) = \int_{\theta} f(D|\theta) f(\theta) \, d\theta
\]

Marginal probability of the data (denominator in Bayes' rule). This is a weighted average of the likelihood, where the weights are provided by the prior distribution.

\[
\Pr(D|M) = \int_{\theta} f(D|\theta, M) f(\theta|M) \, d\theta
\]

Often left out is the fact that we are also conditioning on M, the model used. \(\Pr(\text{DIM}_1)\) is comparable to \(\Pr(\text{DIM}_2)\) and thus the marginal probability of the data can be used to compare the average fit of different models as long as the data \(D\) is the same.
Bayes Factor: 1-param. model

Average likelihood = \( \left( \frac{1}{2} \right) L_0 \)
Bayes Factor: 2-param. model

Average likelihood = \( \left( \frac{1}{2} \right)^2 L_1 \)
Bayes Factor is ratio of marginal model likelihoods

1-parameter model $M_0$: $(\frac{1}{2}) L_0$
2-parameter model $M_1$: $(\frac{1}{4}) L_1$

$\text{BF}_{01} = \frac{(1/2)L_0}{(1/4)L_1} = \frac{2L_0}{L_1}$

$\text{BF}_{01}$ is Bayes Factor in favor of model $M_0$ against model $M_1$:
- if $\text{BF}_{01} > 1$, model $M_0$ wins
- if $\text{BF}_{01} < 1$, model $M_1$ wins

In this case, $L_1$ would need to be twice as great as $L_0$ in order for model $M_1$ to win.

Notes about $\text{BF}$:
• automatically penalizes model for extra dimensions (parameters)
• severity of penalty depends on priors (under control of investigator, unlike AIC, BIC, LRT, etc., which assess a constant penalty for each additional parameter)

Recent work on Bayes factors with respect to phylogenetics:
Vague priors penalize model more

Model 1 better than Model 2 according to BF

Vague prior in Model 2 simply adds more area in which likelihood is small, bringing down the average likelihood of the model.
Something closer to reality

• Example:
  – Compare JC69 vs. K80 models
  – Parameters:
    • $\nu$ is edge length (expected no. substitutions/site)
      – free in both JC69 and K80 models
    • $\kappa$ is transition/transversion rate ratio
      – free in K80, set to 1.0 in JC69

X ———— Y
Likelihood Surface when K80 true

sequence length = 1000 sites
true branch length = 0.15
true kappa = 4.0

K80 wins
Likelihood Surface when JC true

sequence length  = 1000 sites
true branch length = 0.15
true kappa = 1.0

K80 model (entire 2d space)

JC69 model (just this 1d line)

JC69 wins

ratio of transition rate to transversion rate
Harmonic mean method

\[ f(D) \approx \frac{n}{\frac{1}{L^{(1)}} + \frac{1}{L^{(2)}} + \cdots + \frac{1}{L^{(n)}}} \]

\((L^{(i)} = \text{Likelihood (not log-likelihood) calculated for the }\)
\(i\text{th sample from the MCMC})

\[ \log BF_{12} = \log f(D|M_1) - \log f(D|M_2) \]

MrBayes’ sump command provides the log of the harmonic mean of the sampled likelihoods for each model you run, so all you need to do is subtract.

This is what we **want**: samples from the prior

This is what we **have**: samples from posterior

Unbiased but **unacceptably high variance**

Both methods converge on the correct marginal likelihood if you have an **infinite number of samples**

Currently no good alternative, but there will be soon (but better accuracy/precision comes at a price)

Harmonic mean corrects for fact that you’ve sampled the posterior

Low variance but **unacceptably high bias** (favors the more complex model)