Some of these slides have been borrowed from Dr. Paul Lewis, Dr. Joe Felsenstein. Thanks!

Paul has many great tools for teaching phylogenetics at his web site:

http://hydrodictyon.eeb.uconn.edu/people/plewis
Markov chain Monte Carlo

- Simulates a walk through parameter/tree space.
- Lets us estimate posterior probabilities for any aspect of the model.
- Relies on the ratio of posterior densities between two points.
\[ R = \frac{Pr(\text{Point}_2|\text{Data})}{Pr(\text{Point}_1|\text{Data})} \]

\[ R = \frac{Pr(\text{Point}_2) L(\text{Point}_2)}{Pr(\text{Data})} \]

\[ R = \frac{Pr(\text{Point}_2) L(\text{Point}_2)}{Pr(\text{Point}_1) L(\text{Point}_1)} \]
**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 that the robot tends to stay near the tops of hills.
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 random number (uniform on 0.0 to 1.0), and that number is less than or equal to $R$. 

Currently at 6.20 m  
Proposed at 5.58 m  
$R = 5.58/6.20 = 0.90$

Currently at 6.20 m  
Proposed at 0.31 m  
$R = 0.31/6.20 = 0.05$

Currently at 1.0 m  
Proposed at 2.3 m  
$R = 2.3/1.0 = 2.3$
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 good sign

Burn-in is over right about here

We started off at a very low point

You can use the program Tracer to easily create this type of 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...

Disadvantage: robot takes smaller steps, more time required to explore the same area

Advantage: robot seldom refuses to take proposed steps
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
Poor mixing

Chain is spending long periods of time “stuck” in one place

Indicates step size too large, and 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.

MCRobot

Windows program download from:
http://www.eeb.uconn.edu/people/plewis/software.php
Metropolis-coupled Markov chain Monte Carlo (MCMCMC, or MC³)

- MC³ 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)

What is a heated chain?

$R$ is the ratio of posterior probability densities.

Instead of using $R$ in the acceptance/rejection decisions, a heated chain uses $R^{1+H}$.

Heating a chain makes the surface it explores flatter.

In MrBayes: $H = \text{“Temperature”} \times (\text{The Chain’s index})$

The cold chain has index 0, and the default temperature is 0.2.
Acceptance Probability for chains with Temp = 0.2

<table>
<thead>
<tr>
<th>$R$</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.2</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
</tr>
<tr>
<td>0.8</td>
<td>0.800</td>
<td>0.830</td>
<td>0.853</td>
<td>0.860</td>
</tr>
<tr>
<td>0.4</td>
<td>0.400</td>
<td>0.466</td>
<td>0.519</td>
<td>0.564</td>
</tr>
<tr>
<td>0.01</td>
<td>0.010</td>
<td>0.021</td>
<td>0.037</td>
<td>0.056</td>
</tr>
</tbody>
</table>

Acceptance Probability for chains with Temp = 0.5

<table>
<thead>
<tr>
<th>$R$</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.8</td>
<td>0.800</td>
<td>0.862</td>
<td>0.894</td>
<td>0.915</td>
</tr>
<tr>
<td>0.4</td>
<td>0.400</td>
<td>0.543</td>
<td>0.633</td>
<td>0.693</td>
</tr>
<tr>
<td>0.01</td>
<td>0.010</td>
<td>0.046</td>
<td>0.100</td>
<td>0.158</td>
</tr>
</tbody>
</table>
Heated chains act as scouts for the cold chain.

- Short steps fall short.
- Longer step suggested by scout.
(the following slides come directly from Paul Lewis’ lecture at the Woods Hole Workshop on Molecular Evolution – thanks, Paul).
So, what’s all this got to do with phylogenetics?

Imagine drawing tree topologies randomly from a bin in which the number of copies of any given topology is proportional to the (marginal) posterior probability of that topology. Approximating the posterior of any particular attribute of tree topologies (e.g. existence of group AC in this case) is simply a matter of counting.
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

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

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)


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 uniform from $\kappa-\delta$ to $\kappa+\delta$

The “step size” of the MCMC robot is defined by $\delta$: a larger $\delta$ 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 distributions

Histogram created from a sample of 1000 $\kappa$ values.

lower = 2.907  
upper = 3.604

95% credible interval

mean = 3.234

IV. Prior distributions
Commonly-used Prior Distributions

- For topologies: discrete Uniform distribution

\[
\begin{array}{cccc}
\text{AB} & \text{BC} & \text{BD} & \text{BE} \\
\text{D} & \text{C} & \text{A} & \text{E} \\
\frac{1}{15} & \frac{1}{15} & \frac{1}{15} & \frac{1}{15} \\
\end{array}
\]
Commonly-used Prior Distributions

- **For proportions (e.g. pinvar):** Beta(a, b) distribution

  - **Beta(10,10):** Peaks at 0.5 if \( a = b \) and both greater than 1.
  - **Beta(1,1):** Flat when \( a = b = 1 \).
  - **Beta(1.2,2):** Leans left if \( a < b \).
  - **Beta(0.8,2):** Mean = \( a/(a+b) \).

The graph illustrates the different shapes of the Beta distributions with varying parameters, showing how they can be used to model proportions.
Commonly-used Prior Distributions

- For base frequencies: **Dirichlet**\((a,b,c,d)\) distribution
  \[ a \rightarrow \pi_A, \ b \rightarrow \pi_C, \ c \rightarrow \pi_G, \ d \rightarrow \pi_T \]

Flat prior:
\[ a = b = c = d = 1 \]

Informative prior:
\[ a = b = c = d = 300 \]

(Thanks to Mark Holder for pointing out to me that a tetrahedron could be used for plotting a 4-dimensional Dirichlet)

(stereo pairs)
Commonly-used Prior Distributions

• For GTR model relative rates: 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 $a=b=c=d=e=f=1$
  - all relative rates nearly equal to each other if $a=b=c=d=e=f$ and large (e.g. 300)
  - to create a vague prior that makes the rate of transitions slightly higher than the rate of transversions, could choose $a=c=d=f=1$ 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)
Common Priors (cont.)

- For other **model parameters and branch lengths**: Gamma\((a,b)\) distribution
  - Exponential\((\lambda)\) equals Gamma\((1, \lambda^{-1})\)
  - Mean of Gamma\((a,b)\) is \(a \times b\)
    - mean of an Exponential\((10)\) distribution is 0.1
  - Variance of a Gamma\((a,b)\) distribution is \(a \times b^2\)
    - variance of an Exponential\((10)\) distribution is 0.01

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.
Priors for model parameters with no upper bound

Exponential(2) = Gamma(1,\(\frac{1}{2}\))

Exponential(0.1) = Gamma(1,10)

Gamma(2,1)

Uniform(0,2)

More About Priors

• Running on empty
• Prior as enemy
• Prior as friend
• Flat vs. informative priors
• Proper vs. improper priors
• Hierarchical models
• Empirical Bayes
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/
More About Priors

• Running on empty
• Prior as enemy
• Prior as friend
• Flat vs. informative priors
• Proper vs. improper priors
• Hierarchical models
• Empirical Bayes
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.
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...
<table>
<thead>
<tr>
<th>Internal branch length prior mean 0.000001</th>
</tr>
</thead>
<tbody>
<tr>
<td>The internal branch length prior is calling the shots now.</td>
</tr>
</tbody>
</table>
More About Priors

• Running on empty
• Prior as enemy
• Prior as friend
• Flat vs. informative priors
• Proper vs. improper priors
• Hierarchical models
• Empirical Bayes
Too many parameters, too little information

H H T T T T H

$3/7 = 0.43$

1.0 1.0 0.0 0.0 0.0 0.0 1.0

7 coins flipped once

1 parameter model behaves well

7 parameter model behaves badly

Under maximum likelihood, parameter values tend to go to extremes if there is too little information. Priors *add information* and can keep models in check.
More About Priors

- Running on empty
- Prior as enemy
- Prior as friend
- Flat vs. informative priors
- Proper vs. improper priors
- Hierarchical models
- Empirical Bayes
Flat prior: posterior proportional to likelihood

\[
\text{posterior} \rightarrow f(\theta | D) = \frac{f(D | \theta) f(\theta)}{f(D)} \quad \text{constant}
\]

\[
\propto f(D | \theta) f(\theta) \quad \text{constant}
\]

\[
\propto f(D | \theta) \quad \text{likelihood}
\]

Under a flat prior, the posterior distribution peaks at the same place as the likelihood function, but:
- posterior mean usually differs from the maximum likelihood estimate
- flat priors are not possible for most parameters
More About Priors

- Running on empty
- Prior as enemy
- Prior as friend
- Flat vs. informative priors
- Proper vs. improper priors
- Hierarchical models
- Empirical Bayes
Proper vs. improper priors

Gamma priors are *proper*: total area = 1.0

Uniform prior from 0 to infinity is *improper*: total area = $\infty$

Such (improper uniform) priors can be **truncated** to make them proper, but choosing the truncation point is arbitrary and can have a disturbingly large effect (see Felsenstein 2004 chapter 18)
More About Priors

- Running on empty
- Prior as enemy
- Prior as friend
- Flat vs. informative priors
- Proper vs. improper priors
- Hierarchical models
- Empirical Bayes
In a non-hierarchical model, all parameters are present in the likelihood function.

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

μ is a hyperparameter governing the mean of the edge length prior

Hyperprior: InverseGamma(mean=1, var=10)

Exponential(mean=μ)

\[
L_k = \frac{1}{4} \left[ \frac{1}{4} + \frac{3}{4} e^{-4\nu_1/3} \right] \left[ \frac{1}{4} + \frac{3}{4} e^{-4\nu_2/3} \right] \left[ \frac{1}{4} - \frac{1}{4} e^{-4\nu_3/3} \right] \left[ \frac{1}{4} - \frac{1}{4} e^{-4\nu_4/3} \right] \left[ \frac{1}{4} + \frac{3}{4} e^{-4\nu_5/3} \right]
\]

For example, see Suchard, Weiss and Sinsheimer. 2001. MBE 18(6): 1001-1013.
Empirical Bayes

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

\[ L_k = \frac{1}{4} \left[ \frac{1}{4} + \frac{3}{4} e^{-4\nu_1/3} \right] \left[ \frac{1}{4} + \frac{3}{4} e^{-4\nu_2/3} \right] \left[ \frac{1}{4} - \frac{1}{4} e^{-4\nu_3/3} \right] \left[ \frac{1}{4} - \frac{1}{4} e^{-4\nu_4/3} \right] \left[ \frac{1}{4} + \frac{3}{4} e^{-4\nu_5/3} \right] \]

This uses some aspects of the data to determine some aspects of the prior, which is not acceptable to purists, who prefer using the hierarchical approach.
V. Bayesian model selection
LRT, AIC and BIC only evaluate *part* of a Bayesian model (i.e. the likelihood)

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.
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. \( \Pr(D|M_1) \) is comparable to \( \Pr(D|M_2) \) 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 $M_1$: $(\frac{1}{4}) L_1$

Bayes Factor favors $M_0$ unless $L_1$ is at least \textit{twice} as large as $L_0$

All other things equal, more complex models are penalized by their extra dimensions
Marginal Likelihood of a Model

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

K80 wins

K80 model (entire 2d space)
JC69 model (just this 1d line)
Marginal Likelihood of a Model

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

JC69 model (just this 1d line)
K80 model (entire 2d space)

JC69 wins
Sample values from the prior. In this case, draw proportion of heads from Uniform(0,1)

Compute likelihood for each point drawn from the prior

Marginal likelihood is the arithmetic mean of the sampled likelihoods

Problem: tends to underestimate marginal likelihood because few draws from the prior will be in the highest part of the likelihood
Harmonic Mean Method

Sample values from the posterior.

Compute likelihood for each point drawn from posterior.

Marginal likelihood is the harmonic mean of the sampled likelihoods.

Problem: tends to overestimate marginal likelihood because few draws from the posterior will be in the lowest part of the likelihood.
Thermodynamic Integration¹

- Special MCMC analysis performed in which the distribution explored slowly changes from posterior to prior
- Produces much more accurate² marginal likelihood estimates:

<table>
<thead>
<tr>
<th>log(marg. like.)</th>
<th>MSE</th>
<th>Method</th>
</tr>
</thead>
<tbody>
<tr>
<td>-167.316</td>
<td>29.62</td>
<td>Harmonic mean</td>
</tr>
<tr>
<td>-172.783</td>
<td>0.01</td>
<td>Thermodynamic Integration</td>
</tr>
<tr>
<td>-172.743</td>
<td>0.00</td>
<td>True value</td>
</tr>
</tbody>
</table>

- More computation needed than for typical Bayesian MCMC analysis


²Work in collaboration with Wangang Xie, Meng Hui Chen, Lynn Kuo and Yu Fan. In this case, model and tree were simple enough that the marginal likelihood could be determined analytically (i.e. the true value is known).
How would we like our phylogenetic inference methods to behave?

Ideally, the methods would return the true tree with strong support for every grouping in the tree.

Why is this perfect performance not possible?

• systematic errors
• sampling errors
What properties are important when choosing between methods? Assessments of support for different aspects of the tree should be:

- interpretable
- reliable
- if we do not feel that the statements of support are always believable, then we may prefer to be conservative

Ways to make Bayesian statements of support more conservative:

- Polytomy prior
- data-size dependent priors
- majority-rule consensus trees
- more complex models, robust Bayesian techniques
Simulating from stars

model tree

inferred trees

“expected” support

\[
\frac{1}{3}
\]

\[
\frac{1}{3}
\]

\[
\frac{1}{3}
\]
## Results of star tree simulations

100,000 sites simulated

<table>
<thead>
<tr>
<th>Tree 1</th>
<th>Tree 2</th>
<th>Tree 3</th>
<th>Tree 1</th>
<th>Tree 2</th>
<th>Tree 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.3029</td>
<td>0.2922</td>
<td><strong>0.4049</strong></td>
<td>0.2990</td>
<td>0.3288</td>
<td><strong>0.3722</strong></td>
</tr>
<tr>
<td><strong>0.4607</strong></td>
<td>0.1362</td>
<td>0.4031</td>
<td>0.3172</td>
<td>0.0464</td>
<td><strong>0.6364</strong></td>
</tr>
<tr>
<td><strong>0.6704</strong></td>
<td>0.0975</td>
<td>0.2321</td>
<td>0.1584</td>
<td><strong>0.7969</strong></td>
<td>0.0447</td>
</tr>
<tr>
<td><strong>0.6120</strong></td>
<td>0.1852</td>
<td>0.2028</td>
<td><strong>0.4625</strong></td>
<td>0.3600</td>
<td>0.1775</td>
</tr>
<tr>
<td><strong>0.3605</strong></td>
<td>0.3570</td>
<td>0.2825</td>
<td><strong>0.7077</strong></td>
<td>0.0881</td>
<td>0.2042</td>
</tr>
<tr>
<td><strong>0.5455</strong></td>
<td>0.2505</td>
<td>0.2040</td>
<td>0.0884</td>
<td>0.0262</td>
<td><strong>0.8854</strong></td>
</tr>
<tr>
<td>0.4253</td>
<td><strong>0.4254</strong></td>
<td>0.1493</td>
<td><strong>0.9551</strong></td>
<td>0.0422</td>
<td>0.0027</td>
</tr>
<tr>
<td>0.1595</td>
<td><strong>0.7465</strong></td>
<td>0.0940</td>
<td>0.1826</td>
<td><strong>0.5511</strong></td>
<td>0.2663</td>
</tr>
<tr>
<td><strong>0.4436</strong></td>
<td>0.1697</td>
<td>0.3867</td>
<td>0.3043</td>
<td><strong>0.4224</strong></td>
<td>0.2733</td>
</tr>
<tr>
<td><strong>0.3994</strong></td>
<td>0.3904</td>
<td>0.2102</td>
<td><strong>0.6559</strong></td>
<td>0.0707</td>
<td>0.2734</td>
</tr>
<tr>
<td>0.1151</td>
<td><strong>0.5912</strong></td>
<td>0.2937</td>
<td>0.0073</td>
<td><strong>0.9892</strong></td>
<td>0.0035</td>
</tr>
<tr>
<td><strong>0.8333</strong></td>
<td>0.0951</td>
<td>0.0716</td>
<td>0.2703</td>
<td><strong>0.4112</strong></td>
<td>0.3185</td>
</tr>
<tr>
<td><strong>0.8317</strong></td>
<td>0.0736</td>
<td>0.0947</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Tree geometry

figure courtesy of Paul Lewis
Coin-flipping analogy

$p$ is the probability of Heads
Uniform prior, no data.
$\Pr(\text{Head-biased}) = 0.5$
Coin-flipping analogy

\[ X = \{3 \text{ Tails, 2 Heads}\} \]
\[ \Pr(\text{Head-biased}) \approx 0.344 \]
Coin-flipping analogy

\[ X = \{246 \text{ Tails, 254 Heads}\} \]

\[ \Pr(\text{Head-biased}) \approx 0.64 \]
Despite the fact that $p = 0.5$:

$$\Pr(\text{Head-biased}|\text{Data}) \sim \text{Uniform}(0, 1)$$

even as the sample size $\to \infty$. 
The nature of the phenomenon

- Polytomies are given 0 prior probability.

- We are asking methods to choose between several *incorrect* answers.

- *Not* a damning flaw in Bayesian analyses (or an indication of a bug in the software).
Behavior of Bayesian inference on trees drawn from the prior

From ?:

Evolutionary process: JC69
Bayesian model: JC69
c = 100
Behavior of Bayesian inference when the inference model is too simple

From ?:

Evolutionary process: GTR+\Gamma
Bayesian model: JC69
c = 100
Creating a more conservative analysis

Allowing for polytomies in Bayesian analyses:

• polytomies express uncertainty

• must place a prior probability on unresolved trees

• new MCMC proposals must be invented (see ?, for details)
Delete Edge Move
Effects of allowing for polytomies

data from ?
Polytomy MCMC Wrap up

• Allowing unresolved trees is one way to make the Bayesian tree inference more conservative

• Even strong priors in favor of polytomies do not give up too much power
Different priors on the internal and external branches

suggested using strong priors that favor short lengths for the internal branches of the tree. This can lower the support for potentially spurious groupings.
Log-Likelihood for 3 trees

Ln(Likelihood) vs. Internal Branch Length

Three curves represent the log-likelihood for each of the three trees:
- Tree 1 (solid line)
- Tree 2 (dashed line)
- Tree 3 (dotted line)