PhyloMath Lecture 10 – April 6, 2004
M. McAloon

We continue here with Bayesian analyses from last lecture, then will end with the Markov Chain Monte Carlo method for approximating prior probabilities:

Simple abstract example of a Bayesian analysis:

Imagine having 3 hypotheses (only 3 possible, no others): We'll call them A, B, C. (D=data). A,B,C are the 3 possible choices for the only model parameter (call this single discrete parameter θ) in this equation:

$$\Pr(A \mid D) = \frac{\Pr(A, D)}{\Pr(A, D) + \Pr(B, D) + \Pr(C, D)} = \Pr(D)$$

where Pr (D) is the unconditional probability of the data. The posterior probability Pr(A|D) is normally not expressed in terms of joint probabilities, as above, but rather in terms of likelihoods (probability of the data conditional on one hypothesis) and prior probabilities (the probability of the condition):

$$= \frac{\Pr(D \mid A)\Pr(A)}{\Pr(D \mid A)\Pr(A) + \Pr(D \mid B)\Pr(B) + \Pr(D \mid C)\Pr(C)}$$

In the numerator, Pr (D|A) is the likelihood of hypothesis A, and Pr (A) is the prior probability of hypothesis A.

Note added by Paul:
Although you often see the likelihood described "likelihood of the data" it is more correct to say "likelihood of the hypothesis" or "likelihood of the parameter". This is because the data are constant with respect to the likelihood, which is a function of the parameter or hypothesis of interest. Two likelihoods scores computed using different data sets are not even comparable. One can think of a set of mutually exclusive hypotheses (e.g. A, B and C) as values of a discrete parameter. This is the viewpoint taken in this example. You can, however, also think of each value of a continuous model parameter as a different hypothesis (one of an infinite number of hypotheses). For example, under the K80 model, you can imagine $\kappa = 1.0$ as being a different hypothesis than $\kappa = 1.0001$.

The denominator, Pr(A)Pr(D|A) + Pr(B)Pr(D|B) + Pr(C)Pr(D|C), is a constant because it is the total probability of the data, marginalized over all possible hypotheses. It is often symbolized Pr(D), and often described as the **marginal likelihood of the model** or the

marginal probability of the data. In more complicated models, this marginalized model likelihood has to take all values of all parameters into consideration (such as the infinite number of possible lengths for each branch, the infinite number of possible k values, the finite (but possibly *very* large number) of possible trees, etc.) An infinite number of parameter values makes it difficult to calculate (even for computers) except for very simple models in which there are closed-form solutions to the integrals involved, so it's good that this term cancels out when using MCMC to approximate the posterior distribution!

Law of total probability: because A, B, and C are the only possibilities for our parameter θ, the marginal probability of the data makes use of the law of total probability, whcih says that you can obtain the unconditional probability of D (that is, Pr(D)) given that you sum over all possible joint probabilities of the form Pr(D,θ).

If we set the prior probabilities for A, B, C all to the same value (this is said to be a **flat** or **uninformative prior**), the posterior probability calculation looks like this:

$$\Pr(A \mid D) = \frac{\Pr(D \mid A)\left(\frac{1}{3}\right)}{\Pr(D \mid A)\left(\frac{1}{3}\right) + (B \mid A)\left(\frac{1}{3}\right) + (C \mid A)\left(\frac{1}{3}\right)}$$

Use of a flat prior in Bayesian analyses results in an analysis that is identical to a maximum likelihood analysis *if the goal is only to find the optimal hypothesis or parameter value*. Note that the 1/3 term can be factored out of all three terms in the denominator above, and then can be cancelled with the 1/3 in the numerator, leaving just

$$\Pr(A \mid D) = \frac{\Pr(D \mid A)}{\Pr(D \mid A) + (B \mid A) + (C \mid A)}$$

It is clear from the above expression that whichever hypothesis (A, B or C) has the largest likelihood will also end up with the largest posterior probability.

Note added by Paul
Bayesians can sometimes successfully argue that maximum likelihood is just a Bayesian analysis using a flat prior. While this is strictly true for a simple problem like the above, involving only discrete choices of hypotheses, the argument becomes more complicated for models (like the K80 model) in which there are parameters that range from 0 to infinity. In thoses cases, it is not possible to apply a completely flat prior.

More compicated but concrete example of a Bayesian analysis:

Note: In PAUP there is a command called "pairdiff" for computing nucleotide-pair frequencies, proportion of sites differing, and apparent transition/transversion ratio. The following shows the results of using the pairdiff command on an rbcL data set containing sequences for corn and tobacco.

corn vs. tobacco

rbcL	A	C	G	T
A	316	11	22	9
C	9	214	9	27
G	25	4	280	11
T	13	29	7	328

n1= 103 transitions
n2= 73 transversions
n3= 1138 no change
n= 1314 base pairs

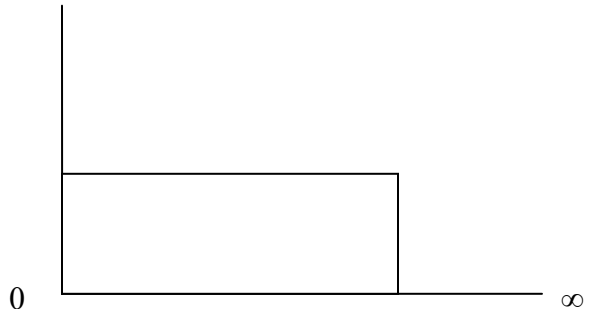
This chart shows all possible substitution types; e.g., for every base A in tobacco, there were 316 A's at the same sequence position in corn. After adding transitions, transversions to sites that did not change we end up with 1314 base pairs.

This model only uses pooled transv/transit/no changes; e.g., it treats C→A and A→C as the same and does not distinguish between the two.

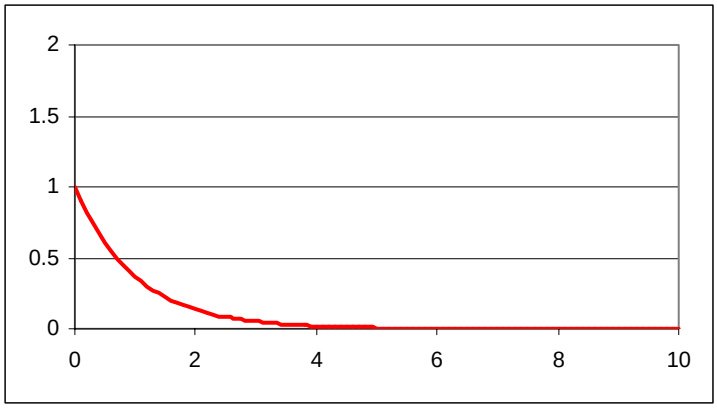
Model characteristics:
Parameters for the K80 model → d = branch length, κ = trs/trv rate ratio
Posterior density $f(d, \kappa \mid n_1, n_2, n_3)$
Note that this time the posterior density is a function of two parameters and thus is a 3-D surface rather than a 2-D curve as we have seen before).

Note added by Paul:
To compute a posterior density, we must specify prior densities for each parameter. WE are talking densities here because the parameters are continuous. In the privious simple example, the prior distributions involved specification of probabilities because the parameter was discrete. The priors chosen in this example are both exponential distributions.

The rationale behind choosing a prior: we cannot use a truly flat prior for a parameter like *d* because that would result in a prior density that has infinite area. Proper densities integrate to 1.0, so we must resort to using something that is not truly uninformative if we are to use a proper prior density. Values of *d* heading way out towards infinity are not really as believable as values closer to 0. We can almost guarantee that within a particular branch there aren't going to be a large number such as 1,000,000 substitutions (this is before we even consider the data). This logic also applies for "κ". If such values were common, we would probably never be able to make sense of any sequence dataset! There are at least two ways to choose a proper prior in this case: MrBayes allows you to choose between a truncated uniform prior:



or an exponential distribution, which places the most prior probability on values that are relatively close to 0:



Note added by Paul:
The truncated uniform prior is not a flat prior (although it is flat out to a point). The exponential prior is a better choice because it *allows* for all possible values of the parameter while still *encouraging* smaller values over really large values.

How to go about calculating the posterior density:
$f(d, \kappa \mid n_1, n_2, n_3)$ = height in "hill space", where "hill space" is the 2-D space defined by the two axes representing *d* and κ where the "hill" (or posterior density surface) resides:

$$f(n_1, n_2, n_3 \mid d, \kappa) = \frac{f(n_1, n_2, n_3 \mid d, \kappa) f(d, \kappa)}{\int_\theta f(n_1, n_2, n_3 \mid d, \kappa)\, f(d, \kappa)\, d\theta}$$

Where the likelihood = $f(n_1, n_2, n_3 \mid d, \kappa)$, the prior = $f(d, \kappa)$, and the denominator integrates over all possible values of d, κ.

Note added by Paul:
The symbol θ is used to indicate the set of all parameters (i.e. d and κ). You could just as easily express the posterior density in terms of two integrals, one "summing" over all possible values of κ, and the other nested integral "summing" over all possible values of d given a particular value of κ:

$$f(n_1, n_2, n_3 \mid d, \kappa) = \frac{f(n_1, n_2, n_3 \mid d, \kappa) f(d, \kappa)}{\int_\kappa \int_d f(n_1, n_2, n_3 \mid d, \kappa) f(d, \kappa)\, dd\, d\kappa}$$

This sort of thing gets very confusing with more complex models with perhaps dozens of parameters, so you often see all the parameters subsumed under one umbrella symbol such as θ.

$f(d, \kappa)$ is shown as a joint prior density because d and κ are not independent… remember that in the K 80 model $d=(\kappa+2)\beta t$ therefore $f(d, \kappa) = f(\kappa) f(d \mid \kappa)$.
Thus, we are assuming an exponential prior for κ, and assuming an independent exponential prior for $d|\kappa$.

Note added by Paul:
In hindsight, I should have parameterized this model differently to avoid the compication described above. It is somewhat strange and overly compicated to assume an exponential prior for $d \mid \kappa$ rather than just d alone. Next time, I will parameterize the model instead in terms of β and κ. The parameter β is independent of κ, so $f(\beta \mid \kappa) = f(\beta)$ and thus the bivariate joint prior can be decomposed into two independent univariate priors, which is eminently more sensible!

$$f(\beta, \kappa) = f(\kappa) f(\beta \mid \kappa) = f(\kappa) f(\beta)$$

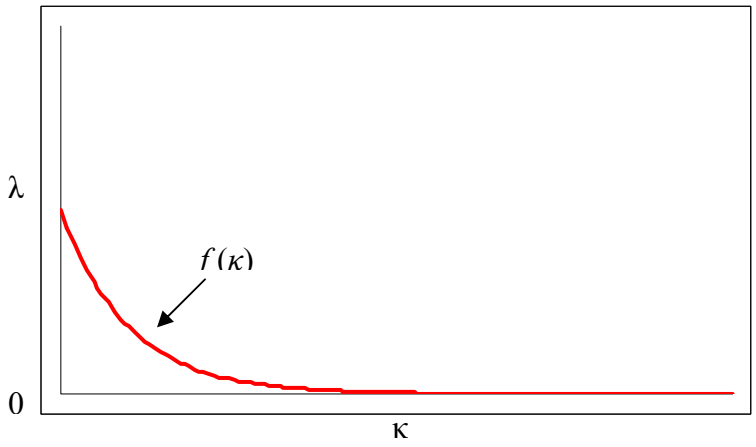
The posterior equals a numerator comprising the likelihood (three terms) and the prior (two terms), divided by the marginal likelihood (C):

$$\frac{[\text{Pr trs}]^{n1}\, [\text{Pr trv}]^{n2}\, [\text{Pr no change}]^{n3}\, (\mu e^{-\mu d})(\lambda e^{-\lambda \kappa})}{C}$$

$$= \left(\frac{1}{C}\right)\left[\left(\frac{1}{4}\right)\left(\frac{1}{4}+\frac{1}{4}e^{-4\beta t}-\frac{1}{2}e^{-4\beta t\left(\frac{k+1}{2}\right)}\right)\right]^{n1}\left[\left(\frac{1}{4}\right)\left(\frac{1}{4}-\frac{1}{4}e^{-4\beta t}\right)\right]^{n2}\left[\left(\frac{1}{4}\right)\left(\frac{1}{4}+\frac{1}{4}e^{-4\beta t}-\frac{1}{2}e^{-4\beta t\left(\frac{k+1}{2}\right)}\right)\right]$$
$$\cdot \left(\mu e^{-\mu d}\right)\left(\lambda e^{-d\kappa}\right)$$

where the first term (raised to the power n_1) is the probability of a transition (e.g., A→G), the second term (raised to n_2) represents the probability of a transversion (e.g. A→T), the third term (raised to n_3) is the probability of no change, the fourth term $\left(\mu e^{-\mu d}\right)$ is the exponential prior density for $d|\kappa$, and the fifth term $\left(\lambda e^{-d\kappa}\right)$ is the exponential prior density for κ.

With increasing κ, $f(\kappa)$ gets smaller. If κ=0, $f(\kappa)$ is at λ, which is as high as it can go.



Larger and larger values of κ make the prior density smaller, and one can always find a value of κ so large that the prior for κ will override anything the data has to say. Increased values of κ thus cause posterior distribution to slope down.

Similar things can be said about d, except that the parameter of this distribution is μ instead of λ.

If you don't want to use default priors (e.g., from MrBayes), you can try different explicit priors, then make an argument for which prior is better.

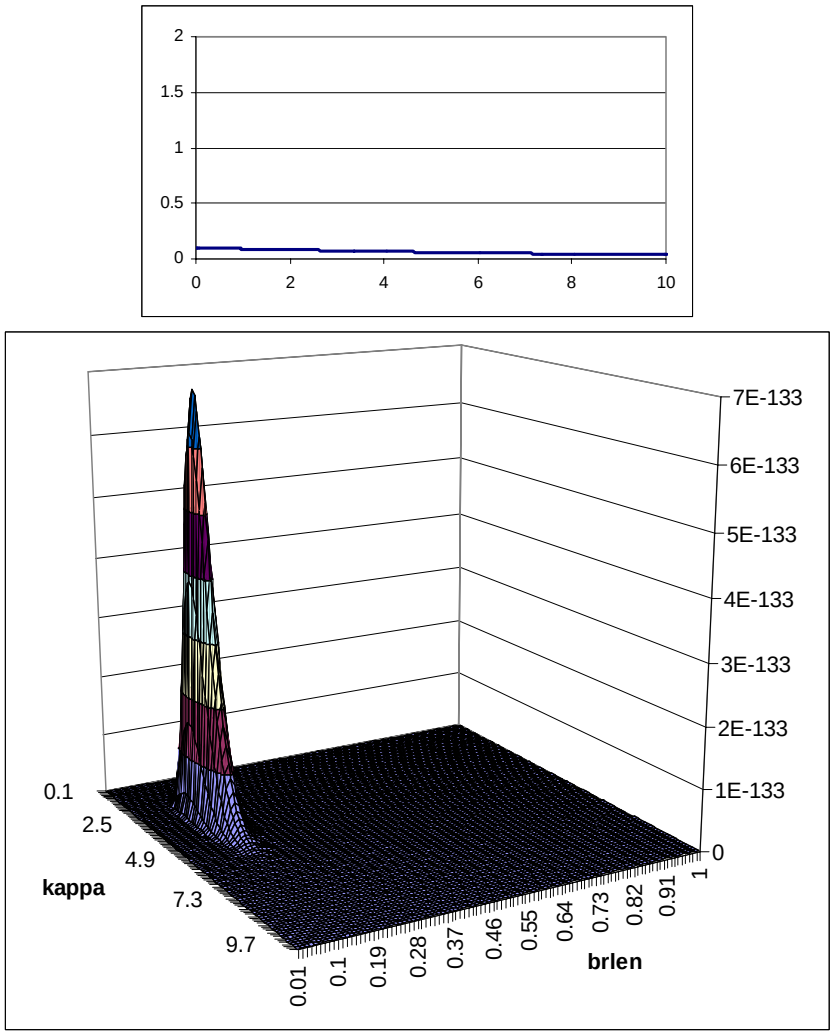
You can use a truncated uniform prior for kappa (κ) where the truncation point is arbitrary. Truncating can make the analysis act strangely. It is better to let all possible values of κ be possible.

When eyeballing the "hill" in 3-D space, we find κ values between 1 and 4.5 account for most of the posterior probability mass. We are more certain about d than κ because the density as viewed from the d axis is relatively less spread out than the posterior density as viewed from the κ axis.

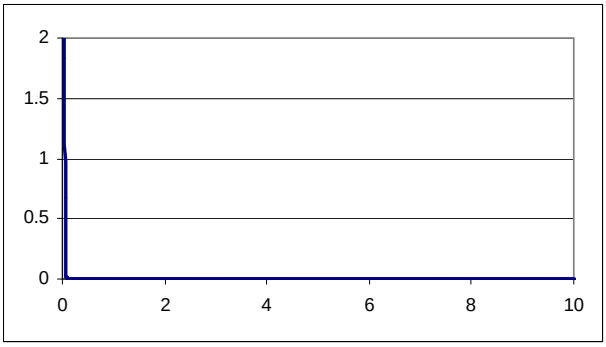
As a general rule, increasing the complexity of models (i.e. adding more independent parameters) increases the uncertainty of any parameter because the posterior density is allowed to spread out into more dimensions.

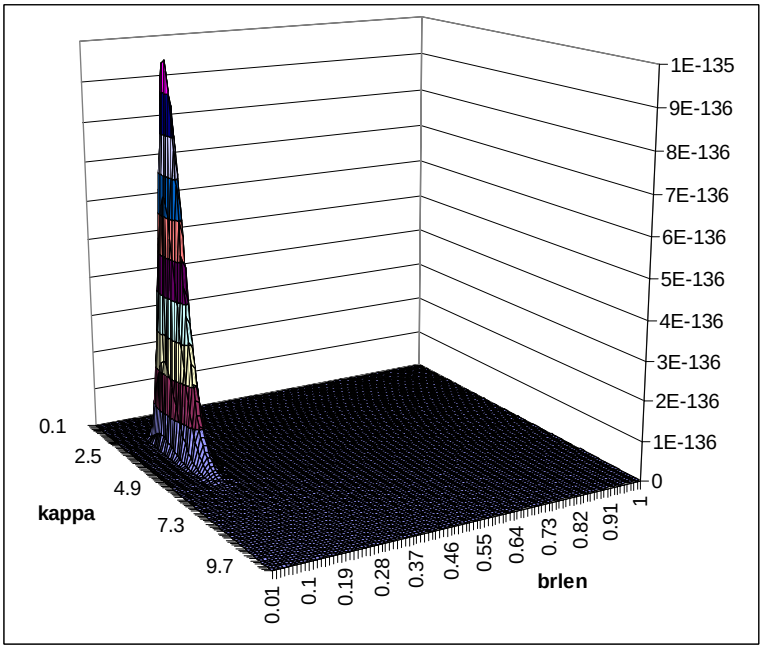
Now we'll look at Paul Lewis' Bayesian Distribution Excel worksheet model:

In the figure below we have set the prior for $d|\kappa$ (branch lengths) to an exponential (0.1) distribution. That is, we have set μ=0.1



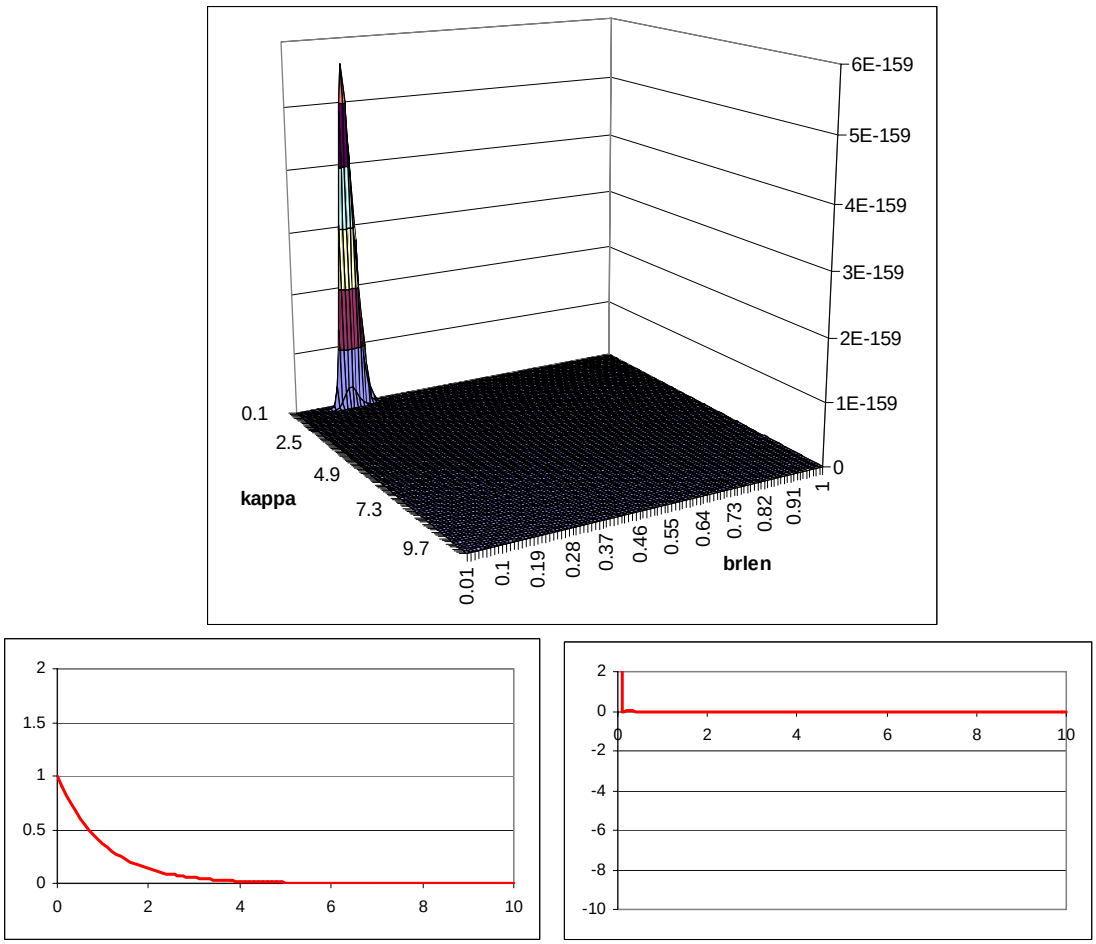
Now if we set the prior for d to, e.g. μ=100 (from the original prior set using μ=0.1), we will not see much difference in the outcome:





Even in this example of a strong prior (μ=100) for branch lengths $d|\kappa$ doesn't affect the shape of the "hill" (or "cone") as much because the data strongly affect the results (the data contains a lot of information about d).

On the other hand, changing the mean of the κ prior results in a shift in posterior probability mass, indicating that a "strong"prior for κ will have more influence on the outcome. Here is the result of changing the κ prior from 1 to .01 (see original 3-D graph above):



The figures directly above show κ with an exponential(1) prior (i.e. λ=1) on the left, and exponential(100) prior (i.e. λ=100) on the right.

Take home message: We should consider playing with priors to see how they influence the posterior density.

Note from Paul:
In practice, experimenting with priors can take a lot of time, so most people (including yours truly) do not do as much of it as we should. Just because most people ignore priors, however, does not mean that determining the sensitivity of the analysis to priors is unimportant!

Note: an example of selecting a realistic prior: we could obtain data and estimate branch lengths on a group closely related to the group in question (e.g. the sister group). We could then use the mean of these branch lengths to set the prior on the group of interest. **Empirical bayesian** analyses involve setting priors based on maximum likelihood estimates using the exact same data used to compute the posterior. In such cases, the priors are generally centered over the maximum likelihood estimates, but are made quite vague so that there is not much of a dependency on the data.

Markov Chain Monte Carlo (MCMC) methods for approximating posterior distribution

For many models, the posterior distribution is not easily calculated exactly; the constant term we looked at earlier is way too complicated, therefore we have to pay a price: approximations. Using simulations, we can approximate the posterior distribution to an abitrary precision at the cost of increased run time.

Metropolis algorithm – first described in 1953
Hastings modification – proposed in 1970
Gelfand and Smith – 1990 paper started the current craze in using simulation algorithms to approximate posterior distributions in Bayesian analyses

Metropolis, N., A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller, and E.Teller. 1953. Equation of state calculations by fast computing machines. J. Chem. Phys. 21:1087-1092.

Hastings, W. K. 1970. Monte Carlo sampling methods using Markov chains and their applications. Biometrika 57:97-109.

Gelfand, A. E., and A. F. M. Smith. 1990. Sampling based approaches to calculating marginal densities. J. Am. Statist. Assoc. 85:398-409.

Simulating a Markov Chain

A Markov Chain is simply a sequence of possible states with certain characteristics, most notably possessing the Markov property.

For example, imagine again our 3 hypotheses A, B, C from the start of this lecture.

We want to be able to generate a sequence of those states such that we can approximate the stationary distribution of the states by simply counting how many times each state occurs in that sequence:

Take sample 1: …B,A,A,A,C,A,A,A,A,A,….
(then thousand steps later) take sample 2…A, A, A, A, B, A, A, A, A….

We're interested in probability distribution of these states:

π_A = .80 (80% of the time we see state A)
π_B = .15
π_C = .05

π_A represents the true posterior probablity of hypothesis A.

Counting states from sample 1:
8/10 bases are A = 80%

1/10 bases are B = 10%
1/10 bases are C = 10%

Counting states from sample 2:

9/10 are A = 90%
1/10 are B = 10%
0/10 are C = 0%

The above samples of the chain reveal similar approximations for the "real" distribution of A, B, C. By using this method we can at least get a good approximation for π_A, π_B, and π_C. We do this to avoid having to calculate the constant (marginal likelihood) we discussed earlier.

Stationarity – probabilities of different states (i.e. the probability distribution) not changing over time.

Another property of Markov chains: statements about the probability of what is going to happen next are not influenced by the previous history of the chain

Consider this Transition Matrix:

	A	B	C
A	P_{AA}	P_{AB}	P_{AC}
B	P_{BA}	P_{BB}	P_{BC}
C	P_{CA}	P_{CA}	P_{CC}

The probability that the chain will be sitting on state B (at time t+1) given that it is sitting on state A at time t is P_{AB}. If this is a Markov chain, then P_{AB} should not depend in any way on the state present at time t-1 or any time before t. This lack of "memory" is called the Markov property.

The probability of ending up in state A after one time period (i.e. time t+1), *regardless* of which state the chain is sitting on at time t can be computed as follows:

$$\pi_A^{(t+1)} = \pi_A^{(t)} P_{AA} + \pi_B^{(t)} P_{BA} + \pi_C^{(t)} P_{CA}$$

If stationarity is assumed, then π_A, π_B and π_C do not change over time so we can drop the references to time, such as (t) and (t+1):

$$\pi_A = \pi_A P_{AA} + \pi_B P_{BA} + \pi_C P_{CA} \qquad \text{(eq. 1)}$$

The goal is to come up with values for P_{ij} such that the above statement is true (as well as similar statements concerning π_B and π_C).

It turns out that if our chain is not only stationary but time reversible, then the above statement is automatically true. Assuming time reversibility means

Pr(start at state ***i*** and end up at ***j***) = Pr(start at state ***j*** and end up at ***i***)

Mathematically, the above can be expressed like this:

$$\pi_i P_{ij} = \pi_j P_{ji}$$

Here's why equation 1 above is guaranteed to be true if time reversibility applies:

$$\begin{aligned}
\pi_i &= \sum_j \pi_j P_{ji} && \text{(this is just eq. 1 stated in general terms)} \\
&= \sum_j \pi_i P_{ij} && \text{(here's where time reversibility is used)} \\
&= \pi_i \sum_j P_{ij} && \text{(only the terms involving j need to stay inside the sum)} \\
&= \pi_i (1) && \text{(the sum of an entire row of the P matrix is 1.0)}
\end{aligned}$$

How do we construct a **P** matrix that guarantees stationarity? If we can construct a P matrix that is time reversible, we get stationarity as a useful byproduct. Why do we want stationarity? We are interested in using the Markov chain to estimated the (posterior) probability of the different states. If our model is correctly formulated, there is just one true posterior probability associated with each state, and thus it would not be very helpful if our Markov chain did not converge to a single value for this probability.

Note added by Paul:
Next time we will show how a very simple definition of the transition matrix P leads to a stationary Markov chain, and how we can simulate such a chain to approximate the stationary distribution of the chain, which also happens to be our posterior distribution of interest.