Bayesian Phylogenetics

Bret Larget

Departments of Botany and of Statistics
University of Wisconsin—Madison

October 1, 2009

Who was Bayes?

- The Reverend Thomas Bayes was born in London in 1702.
- He was the son of one of the first Noncomformist ministers to be ordained in England.
- He became a Presbyterian minister in the late 1720s, but was well known for his studies of mathematics.
- He was elected a Fellow of the Royal Society of London in 1742.
- He died in 1761 before his works were published.

What is Bayes' Theorem?

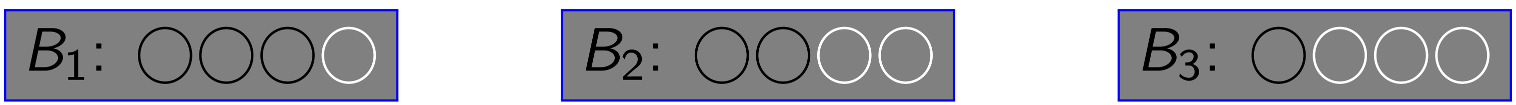
- Bayes' Theorem explains how to calculate inverse probabilities.
- For example, suppose that Box B_1 contains four balls, three of which are black and one of which is white.
- Box B_2 has four balls, two of which are black and two of which are white.
- Box B_3 has four balls, one of which is black and three of which are white.

B_1: ○○○○ B_2: ○○○○ B_3: ○○○○

- If a ball is chosen *uniformly at random* from Box B_1, there is a $3/4$ chance that it is black.
- But if a black ball is drawn, how likely is it that it came from Box B_1?

What is Bayes' Theorem?

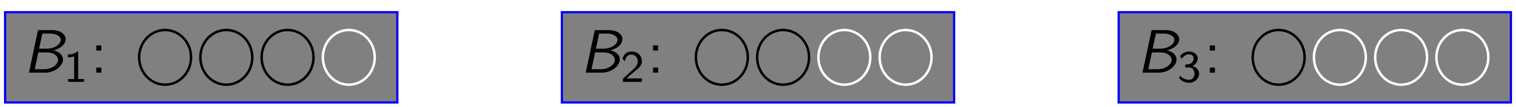
- Bayes' Theorem explains how to calculate inverse probabilities.
- For example, suppose that Box B_1 contains four balls, three of which are black and one of which is white.
- Box B_2 has four balls, two of which are black and two of which are white.
- Box B_3 has four balls, one of which is black and three of which are white.



- If a ball is chosen *uniformly at random* from Box B_1, there is a $3/4$ chance that it is black.
- But if a black ball is drawn, how likely is it that it came from Box B_1?

What is Bayes' Theorem?

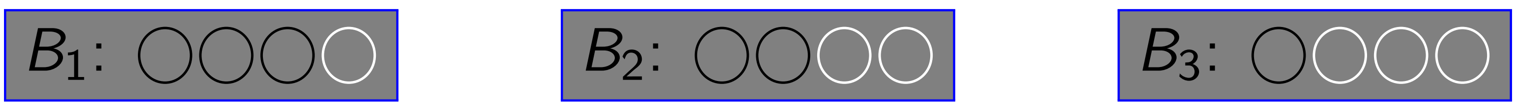
- Bayes' Theorem explains how to calculate inverse probabilities.
- For example, suppose that Box B_1 contains four balls, three of which are black and one of which is white.
- Box B_2 has four balls, two of which are black and two of which are white.
- Box B_3 has four balls, one of which is black and three of which are white.



- If a ball is chosen *uniformly at random* from Box B_1, there is a $3/4$ chance that it is black.
- But if a black ball is drawn, how likely is it that it came from Box B_1?

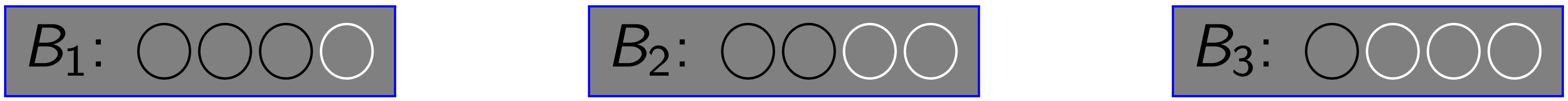
What is Bayes' Theorem?

- Bayes' Theorem explains how to calculate inverse probabilities.
- For example, suppose that Box B_1 contains four balls, three of which are black and one of which is white.
- Box B_2 has four balls, two of which are black and two of which are white.
- Box B_3 has four balls, one of which is black and three of which are white.



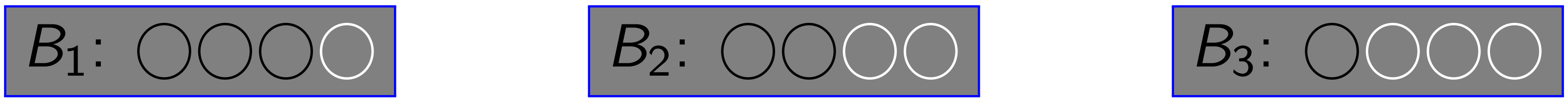
- If a ball is chosen *uniformly at random* from Box B_1, there is a $3/4$ chance that it is black.
- But if a black ball is drawn, how likely is it that it came from Box B_1?

What is Bayes' Theorem?



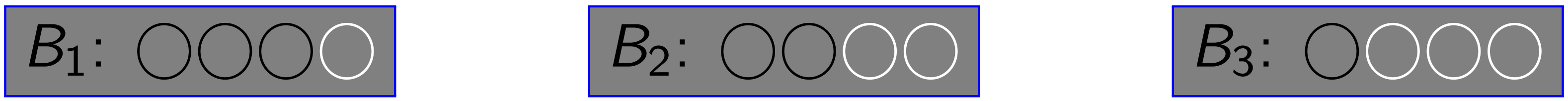
- If a black ball is drawn, how likely is it that it came from Box B_1?
- To answer this question, we need to have prespecified probabilities of which box we pick to draw the ball from.
- The answer will be different if we believe *a priori* that Box B_1 is 10% likely to be the chosen box than if we believe that all three boxes are equally likely.
- Do the problem with a probability tree.

What is Bayes' Theorem?



- If a black ball is drawn, how likely is it that it came from Box B_1?
- To answer this question, we need to have prespecified probabilities of which box we pick to draw the ball from.
- The answer will be different if we believe *a priori* that Box B_1 is 10% likely to be the chosen box than if we believe that all three boxes are equally likely.
- Do the problem with a probability tree.

What is Bayes' Theorem?



- If a black ball is drawn, how likely is it that it came from Box B_1?
- To answer this question, we need to have prespecified probabilities of which box we pick to draw the ball from.
- The answer will be different if we believe *a priori* that Box B_1 is 10% likely to be the chosen box than if we believe that all three boxes are equally likely.
- Do the problem with a probability tree.

Bayes' Theorem

- Bayes' Theorem states that if a complete list of mutually exclusive events $B_1, B_2, \ldots$ have prior probabilities $\mathrm{P}(B_1), \mathrm{P}(B_2), \ldots$, and if the *likelihood* of the event A given event B_i is $\mathrm{P}(A \mid B_i)$ for each i, then

$$\mathrm{P}(B_i \mid A) = \frac{\mathrm{P}(A \mid B_i)\mathrm{P}(B_i)}{\sum_j \mathrm{P}(A \mid B_j)\mathrm{P}(B_j)}$$

- The *posterior probability* of B_i given A, written $\mathrm{P}(B_i \mid A)$, is proportional to the product of the *likelihood* $\mathrm{P}(A \mid B_i)$ and the *prior probability* $\mathrm{P}(B_i)$ where the normalizing constant $\mathrm{P}(A) = \sum_j \mathrm{P}(A \mid B_j)\mathrm{P}(B_j)$ is the prior probability of A.

Bayes' Theorem

- Bayes' Theorem states that if a complete list of mutually exclusive events $B_1, B_2, \ldots$ have prior probabilities $\mathrm{P}(B_1), \mathrm{P}(B_2), \ldots$, and if the *likelihood* of the event A given event B_i is $\mathrm{P}(A \mid B_i)$ for each i, then

$$\mathrm{P}(B_i \mid A) = \frac{\mathrm{P}(A \mid B_i)\mathrm{P}(B_i)}{\sum_j \mathrm{P}(A \mid B_j)\mathrm{P}(B_j)}$$

- The *posterior probability* of B_i given A, written $\mathrm{P}(B_i \mid A)$, is proportional to the product of the *likelihood* $\mathrm{P}(A \mid B_i)$ and the *prior probability* $\mathrm{P}(B_i)$ where the normalizing constant $\mathrm{P}(A) = \sum_j \mathrm{P}(A \mid B_j)\mathrm{P}(B_j)$ is the prior probability of A.

Connection to Phylogeny

- In a Bayesian approach to phylogenetics, the boxes are like different tree topologies, only one of which is right.
- The colored balls are like site patterns, except that there are many more than two varieties and we are able to observe multiple independent draws from each box.
- Things are further complicated in that additional parameters such as branch lengths and likelihood model parameters affect the likelihood, but are also unknown.

Connection to Phylogeny

- In a Bayesian approach to phylogenetics, the boxes are like different tree topologies, only one of which is right.
- The colored balls are like site patterns, except that there are many more than two varieties and we are able to observe multiple independent draws from each box.
- Things are further complicated in that additional parameters such as branch lengths and likelihood model parameters affect the likelihood, but are also unknown.

Connection to Phylogeny

- In a Bayesian approach to phylogenetics, the boxes are like different tree topologies, only one of which is right.
- The colored balls are like site patterns, except that there are many more than two varieties and we are able to observe multiple independent draws from each box.
- Things are further complicated in that additional parameters such as branch lengths and likelihood model parameters affect the likelihood, but are also unknown.

Prior and Posterior Distributions

- A *prior distribution* is a probability distribution on parameters *before* any data is observed.
- A *posterior distribution* is a probability distribution on parameters *after* data is observed.

Prior and Posterior Distributions

- A *prior distribution* is a probability distribution on parameters *before* any data is observed.
- A *posterior distribution* is a probability distribution on parameters *after* data is observed.

Bayesian Methods vs. Maximum Likelihood

	Maximum Likelihood	Bayesian
Probability	Only defined in the context of long-run relative frequencies	Describes everything that is uncertain
Parameters	Fixed and Unknown	Random
Nuisance Parameters	Optimize them	Average over them
Testing	p-values	Bayes' factors
Nature of Method	Objective	Subjective

Bayesian Phylogenetic Methods

- Let's say we want to find the posterior probability of a clade.
- Then we are interested in computing

$$P(\text{clade} \mid \text{data}) = \sum_{\text{tree with clade}} P(\text{tree} \mid \text{data}) = \sum_{\text{tree with clade}} \frac{P(\text{data} \mid \text{tree}) P(\text{tree})}{P(\text{data})}$$

- But we need to know the parameters including branch lengths (params) to compute the likelihood.

$$\sum_{\text{tree with clade}} P(\text{data} \mid \text{tree}) P(\text{tree})$$

$$= \sum_{\text{tree with clade}} \int P(\text{data}, \text{params} \mid \text{tree}) P(\text{tree}) d\text{params}$$

$$= \sum_{\text{tree with clade}} P(\text{tree}) \int P(\text{data} \mid \text{params}, \text{tree}) P(\text{params} \mid \text{tree}) d\text{params}$$

Bayesian Phylogenetic Methods

- Let's say we want to find the posterior probability of a clade.
- Then we are interested in computing

$$\begin{aligned}
\mathrm{P(clade \mid data)} &= \sum_{\text{tree with clade}} \mathrm{P(tree \mid data)} \\
&= \sum_{\text{tree with clade}} \frac{\mathrm{P(data \mid tree) P(tree)}}{\mathrm{P(data)}}
\end{aligned}$$

- But we need to know the parameters including branch lengths (params) to compute the likelihood.

$$\begin{aligned}
&\sum_{\text{tree with clade}} \mathrm{P(data \mid tree) P(tree)} \\
&= \sum_{\text{tree with clade}} \int \mathrm{P(data, params \mid tree) P(tree) d params} \\
&= \sum_{\text{tree with clade}} \mathrm{P(tree)} \int \mathrm{P(data \mid params, tree) P(params \mid tree) d params}
\end{aligned}$$

Bayesian Phylogenetic Methods

- Let's say we want to find the posterior probability of a clade.
- Then we are interested in computing

$$\begin{aligned} \mathrm{P}(\text{clade} \mid \text{data}) &= \sum_{\text{tree with clade}} \mathrm{P}(\text{tree} \mid \text{data}) \\ &= \sum_{\text{tree with clade}} \frac{\mathrm{P}(\text{data} \mid \text{tree})\mathrm{P}(\text{tree})}{\mathrm{P}(\text{data})} \end{aligned}$$

- But we need to know the parameters including branch lengths (params) to compute the likelihood.

$$\begin{aligned} & \sum_{\text{tree with clade}} \mathrm{P}(\text{data} \mid \text{tree})\mathrm{P}(\text{tree}) \\ = & \sum_{\text{tree with clade}} \int \mathrm{P}(\text{data}, \text{params} \mid \text{tree})\mathrm{P}(\text{tree})\,d\text{params} \\ = & \sum_{\text{tree with clade}} \mathrm{P}(\text{tree}) \int \mathrm{P}(\text{data} \mid \text{params}, \text{tree})\mathrm{P}(\text{params} \mid \text{tree})\,d\text{params} \end{aligned}$$

Bayesian Phylogenetic Methods

- So, we need to compute:

$$\frac{\sum_{\text{tree with clade}} \mathrm{P}(\text{tree}) \int \mathrm{P}(\text{data} \mid \text{params}, \text{tree}) \mathrm{P}(\text{params} \mid \text{tree}) \mathrm{d}\text{params}}{\mathrm{P}(\text{data})}$$

- However, P(data) is generally not computable.
- Solution? Markov chain Monte Carlo.

Bayesian Phylogenetic Methods

- So, we need to compute:

$$\frac{\sum_{\text{tree with clade}} \mathsf{P}(\text{tree}) \int \mathsf{P}(\text{data} \mid \text{params}, \text{tree}) \mathsf{P}(\text{params} \mid \text{tree}) \mathrm{d}\text{params}}{\mathsf{P}(\text{data})}$$

- However, P(data) is generally not computable.
- Solution? Markov chain Monte Carlo.

Bayesian Phylogenetic Methods

- So, we need to compute:

$$\frac{\sum_{\text{tree with clade}} \mathrm{P}(\text{tree}) \int \mathrm{P}(\text{data} \mid \text{params}, \text{tree}) \mathrm{P}(\text{params} \mid \text{tree}) \mathrm{d}\text{params}}{\mathrm{P}(\text{data})}$$

- However, P(data) is generally not computable.
- Solution? Markov chain Monte Carlo.

Metropolis-Hastings Example

- Assume a Jukes-Cantor likelihood model for two species where we observe 50 sites, 9 of which differ.
- The likelihood for the distance d is

$$L(d) = \left(\frac{1}{4}\right)^{50} \times \left(\frac{1}{4} - \frac{1}{4}\mathrm{e}^{-\frac{4}{3}d}\right)^{9} \times \left(\frac{1}{4} + \frac{3}{4}\mathrm{e}^{-\frac{4}{3}d}\right)^{41}$$

- Assume a prior for d with the form

$$p(d) = \frac{\lambda}{(1+\lambda d)^2}, \quad d > 0$$

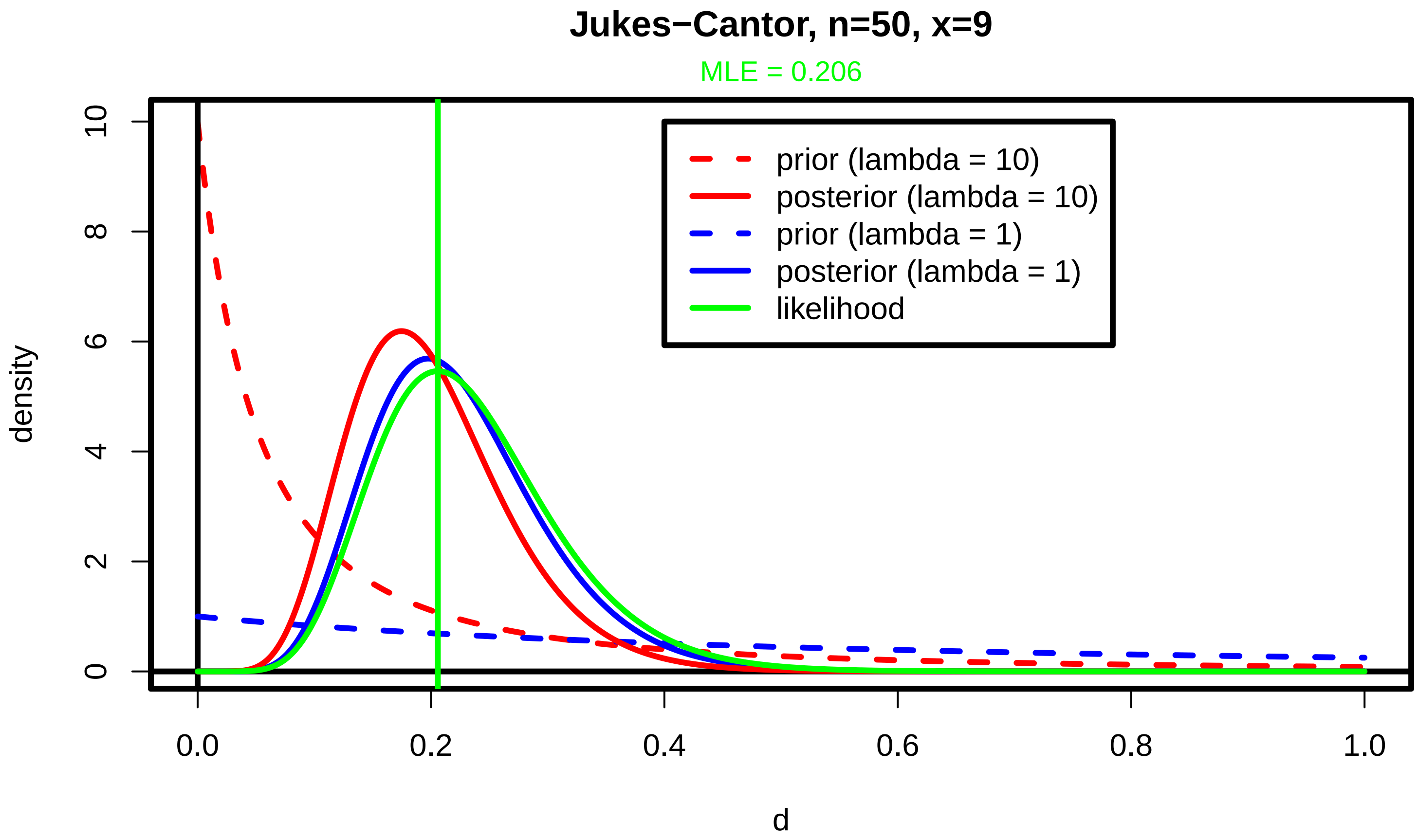
  where $\lambda > 0$ is a parameter.
- This density is what you get if you take the ratio of two independent exponential random variables, one with parameter λ and one with parameter 1.
- The median is $1/\lambda$, but the mean is $+\infty$.

Example

- An exact expression for the posterior density of d is

$$p(d \mid x) = \frac{\left(\frac{\lambda}{(1+\lambda d)^2}\right)\left(\left(\frac{1}{4}\right)^{50}\left(\frac{1}{4}-\frac{1}{4}\mathrm{e}^{-\frac{4}{3}d}\right)^{9}\left(\frac{1}{4}+\frac{3}{4}\mathrm{e}^{-\frac{4}{3}d}\right)^{41}\right)}{\int_0^\infty \left(\frac{\lambda}{(1+\lambda d)^2}\right)\left(\left(\frac{1}{4}\right)^{50}\left(\frac{1}{4}-\frac{1}{4}\mathrm{e}^{-\frac{4}{3}d}\right)^{9}\left(\frac{1}{4}+\frac{3}{4}\mathrm{e}^{-\frac{4}{3}d}\right)^{41}\right)\mathrm{d}d}$$

Graph


Jukes−Cantor, n=50, x=9
MLE = 0.206
prior (lambda = 10)
posterior (lambda = 10)
prior (lambda = 1)
posterior (lambda = 1)
likelihood
density
0
2
4
6
8
10
0.0
0.2
0.4
0.6
0.8
1.0
d