"PRINCIPLES OF PHYLOGENETICS: ECOLOGY AND EVOLUTION"

Integrative Biology 200 Spring 2014

University of California, Berkeley T.L. Grzymala

Updated from N. Matzke & M. Landis

Feb. 19, 2014. Phylogenetic Trees V: Bayesian Inference

I. Introduction

Bayesian Inference (BI) represents one of the three optimality criteria that you can employ for phylogenetic analysis. The other two that we have discussed previously are maximum parsimony (MP) and maximum likelihood (ML).

Except for basic probability, essentially all the statistics that any of you learned in high school and college was frequentist (without saying so). So for most people, frequentist statistics – ideas like chi-squared tests, t-tests, regression, ANOVA, and testing of null hypotheses – simply is "statistics".

The frequentist asks, "What is the likelihood of the data given a hypothesis?" That is, the frequentist says there exists a true parameter, and a true model, and the data are generated through repeated trials.

The Bayesian asks, "What is the probability of a hypothesis given the data?" Bayesian analysis affords a very natural interpretation of the probability of one hypothesis versus another. This is the main attraction of Bayesian analysis, for which frequentist statistics have no equivalent. The null hypothesis becomes just one of all hypotheses that may be explored.

Additional discussion on statistical schools of thought: http://videolectures.net/mlss09uk_jordan_bfway/

Bayesianism is actually much older than Frequentism, dating back at least to the 1700s. The name comes from the Reverend Thomas Bayes (1702–1761), who proposed a special case of what came to be called "Bayes' theorem". This theorem is easiest to understand by starting with basic probability and conditional probability.

II. Basic Probability and Bayes Theorem

Basic Probability

Let's first remember some basic probability.

- P(E) = P(event) = "Probability than an event occurs in a trial"

Often writers talk about the P(data) or P(observations) instead of P(event).

- Probabilities of exclusive events must sum to 1, so P(E) + P(not E) = 1

Discussion Questions:

- What is P(heads)?
- What is P(rolling a 1) = P(event = 1) = P(1)?

Conditional Probability

In reality, to answer the questions above, we need some model or hypothesis before we can calculate the probability. This is:

- P(event given some model/hypothesis) = P(event | hypothesis) = P(E | H)
- "model" and "hypothesis" get used interchangeably

E.g., the probability of getting a 1 *on a 6-sided fair die* is

- P(event = 1 | "6-sided fair die") = 1/6, or
- P(E|H) = 1/6, where E="rolling a 1" and H="die is six-sided and fair"

What is the probability of rolling a 1 if the die is randomly picked from 2 dice, where 1 die is 6-sided and fair, and 1 die is 6-sided and all 1s?

- P(event=1 | "fair die") / P("fair die") + P(event=1 | "die w/ all 1s") / P("die w/all ones)
- = P(E=1 | H1) / P(H1) + P(E=1 | H2) / P(H2), where H1 = fair die, H2 = all ones
- = P(E|H1) / P(H1) + P(E|H2) / P(H2)
- = (1/6) / (1/2) + (6/6) / (1/2)
- = 1/12 + 1/2 = 7/12

P(E|H) is a *conditional probability*, i.e. the probability of E given H.

The above example, or thinking of probability in terms of proportions gives us Kolmogorov's (1950) definition of conditional probability (Sober 2008, p. 9):

- P(E | H) = P(E & H) / P(H)

Both E and H, while we call them "events" and "hypotheses", are really both just propositions. Randomly rolling a "1" is no different than randomly picking the fair or unfair die. So E and H can be switched:

- P(H | E) = P(H & E) / P(E)

Since P(H & E) and P(E & H) are the same thing, we can say something interesting:

- P(E & H) = P(E | H) P(H)
- P(H & E) = P(H | E) P(E)

So,

- P(H | E) P(E) = P(E | H) P(H)
- P(H | E) = P(E | H) P(H) / P(E)

Bayes' theorem

This is the standard version of Bayes' theorem. Let's write the same thing in a few different ways:

$$P(H \mid E) = \frac{P(E \mid H)\ P(H)}{P(E)}$$

$$P(\text{hypothesis} \mid \text{event}) = \frac{P(\text{event} \mid \text{hypothesis})\ P(\text{hypothesis})}{P(\text{event})}$$

$$P(\text{model} \mid \text{data}) = \frac{P(\text{data} \mid \text{model})\ P(\text{model})}{P(\text{data})}$$

$$\text{Posterior probability} = \frac{\text{Likelihood} * \text{Prior probability of the model}}{\text{Unconditional probability of the data}}$$

Notes:

- *Prior probability* is probability of the model, before you look at the data
- *Posterior probability* is the probability of the model, after adding the data
- The *Likelihood* is the *probability that the model confers on the data*. Keep in mind that it is probability *of the data*, *not of the model,* although one might prefer a model if it gives the observed data a higher likelihood than another model.
- The *Unconditional probability of the data* is the probability of the data summed over all possible conditions, i.e. an integral.
    - If we think of probability as proportions, then it makes sense that we would need to *normalize* the numerator of Bayes' theorem so that the posterior probability represents the probability (out of a maximum of 1) of the model.
    - Also known as the "nasty normalizing constant"
    - The integral that gives P(data) is $P(\text{data}) = \int P(\text{data} \mid \text{model})\ P(\text{model})\ d(\text{model})$
    - This integral is very often impossible, except in simple cases, or certain families of distributions

Discrete probabilities versus probability densities

One thing that is confusing to introductory students is that Bayes' theorem is typically introduced using simple problems with discrete probabilities. In real life, often we are trying to estimate continuous parameters like branch lengths and substitution rates. Here, the prior, likelihood, and posterior are represented by *continuous functions*. E.g. the probability density functions for *p*, the proportion of time a coin will turn up heads (where 0.5 = fair coin) is shown in Sober (2008), p. 22:

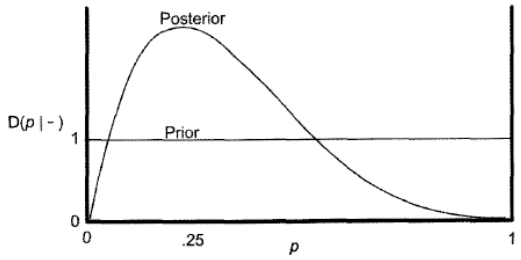


Figure 1.3 A flat prior density distribution for *p* and the non-flat posterior density occasioned by observing one head in four tosses. The prior expected value of *p* is 0.5; given this prior, the posterior expected value of *p* is 0.33.

Under this "flat prior", your initial guess is that *p* has an equal chance of being any value. After observing 1 head in 4 tosses, your posterior reflects that observed data.

The fact that probability densities are represented by functions means that employing Bayes' theorem involves multiplying, dividing, and integrating functions, which can get complex, although there are a number of useful reference works on the web on the relationships between statistical distributions.

Bayesianism: Pros and Cons

The debates over Bayesianism vs. other approaches are fairly epic, but here is a short summary:

Pros:

1. Your beliefs both before and after looking at the new data are explicit, available for all others to judge for themselves.
2. Posterior distributions are better than "point estimates" – i.e., instead of a (point) mean estimate of heights, with a standard deviation, which *assumes* some distribution (e.g. the normal distribution), posteriors can take any shape and are not necessarily controlled by some theoretical assumption.
    - This is why Bayesians talk about *credibility* intervals rather than *confidence* intervals.
3. Bayesian methods can be very flexible, taking into account quite complex models and datasets.
4. The interpretation of posterior probabilities seems fairly obvious and intuitive.

Cons:

1. Getting the nasty normalization constant can be very hard.
    - This is (or at least has been) a practical barrier, not necessarily a philosophical problem.
    - However, this has been ameliorated somewhat by:
        i. Theoretical work, e.g. conjugate priors are known in some situations – a conjugate prior for a certain likelihood function produces a posterior with the same distribution as the prior
        ii. Numerical integration can be attempted when exact mathematical integration is impossible; e.g. MCMC sampling approaches

2. People don't like it because they think all statistics is frequentist.
    - This is a sociological statement, not necessarily an argument, although it is a reason that you need to know the pros and cons of the different approaches, and be able to provide an argument.
    - Scientists tend to be practical, and will go with whatever works for their problem and data.

3. The biggie: how do you choose a prior? Isn't that arbitrary?
    - One response is that everyone is operating under some prior belief, whether or not they admit it, and that it is best to be explicit about it.
    - There are different schools of thought within Bayesianism about how to obtain priors, e.g.

- "Objective Bayesians" try to come up with "unbiased" priors that are maximally agnostic about what the true value is. E.g. Laplace justified the use of "flat priors" with the Principle of Indifference.
    1. The Principle of Indifference is flawed. Sober (2008): One might think a reasonable prior on the existence of God is 0.5 – 50/50 chance he exists or not! But there are other options, e.g. what about Zeus?
    2. However, uniform, flat, priors are not always truly agnostic, e.g. watch out for:
        - The limits on the uniform distribution (ranges from 0 to 10? 0 to infinity?)
        - Depending on the shape of the likelihood curve, a flat prior may actually be putting a lot of weight in an unusual place.
    3. Theoretical work indicates that sometimes you can find a "reference prior", which is a prior that has the minimum theoretical impact on the posterior compared to the data. However, reference priors are not necessarily conjugate or otherwise convenient
    4. Hyperpriors – make the choice of prior itself a variable. (Huelsenbeck: "But the madness has got to stop somewhere!")
- Subjective Bayesians: they love putting prior knowledge into the prior, that is the whole point of having a prior.
    1. E.g., a statistician might consult domain experts to get a sense of what a reasonable prior is for a problem. One might even conduct a formal survey of experts. See e.g. Huelsenbeck et al. (2002), Systematic Biology, p. 678:

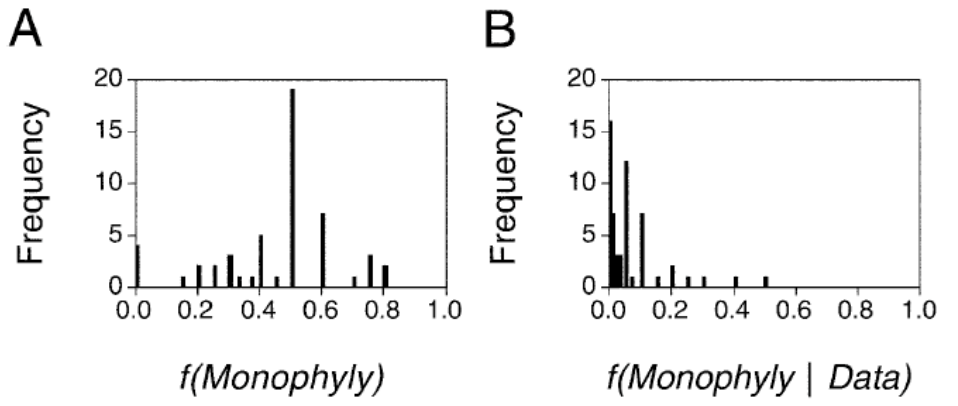


FIGURE 1. Frequency histograms of the responses concerning *Ipomoea* monophyly. (A) Prior beliefs. (B) Updated beliefs.

2. Or, there are other criteria, e.g. convenience and calculation speed are reasons to prefer conjugate priors

- In general, Bayesians hope/expect that the data will "swamp the prior"

III. Bayesian phylogenetics

We have expended a lot of time and effort going through the background of statistical schools of thought. This general background is important as various complicated debates among phylogeneticists often turn out to be expressions of the fundamental debate between different statistical schools of thought, whether this is realized or not. And the Bayesian phylogenetic method makes a lot more sense if you have been introduced to Bayes' theorem first.

The goal of Bayesian phylogenetics

The goal is to find the posterior probability density of a hypothesis/model/model parameters of interest (e.g. a phylogenetic tree, a clade, a substitution rate, etc.):

$$P(\text{tree, parameters} \mid \text{data}) = P(\text{data} \mid \text{tree, parameters})\, P(\text{tree, parameters}) / P(\text{data})$$

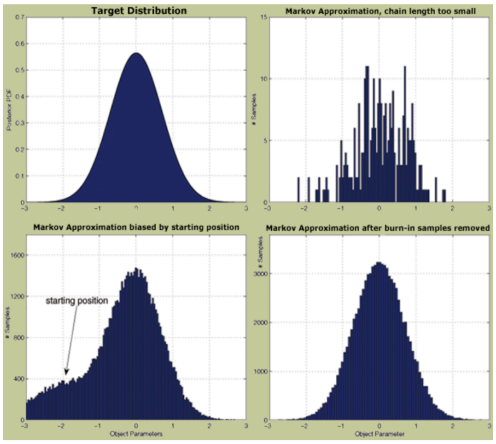
tree = tree topology and branch lengths
parameters = parameters of substitution model (e.g. GTR + I + gamma), other parameters
data = sequence alignment or character matrix

Sampling the posterior distribution with Markov Chain, Monte Carlo (MCMC) method

Monte Carlo: stochastic draws from distributions
Markov Chain: the values of parameters are explored in a series of steps (a "chain")

We're interested in exploring the posterior distribution of the trees and parameters given our data and model. One would like to directly calculate the likelihood for every single value for every single parameter, but this is an impossible task. MCMC is a statistical method to approximate a probability distribution. The algorithm explores the distribution step-by-step, preferring moves into regions of higher probability to those of lower probability. At intervals, MCMC records the current parameter values that correspond to a likelihood. After many steps (eg. millions), the MCMC will have spent time in each of the probability distribution proportional to its posterior probability.

MCMC decides how to move by proposing a new parameter value to explore. If MCMC prefers to explore regions of high probability, you may be worried MCMC would get stuck on local likelihood optima. To avoid this problem, The Metropolis-Hastings (MH) algorithm is used, which allows MCMC to move downhill instead of only climbing uphill.



Upper left: The "true" distribution.

Upper right: Too few MCMC samples

Lower left: Many MCMC samples including burn-in

Lower right: Many MCMC samples excluding burn-in, approximating the "true" distribution

[From M. Landis IB200a lecture 2012]

Details, sometimes problematic:

- There is a bit of an art to thoroughly exploring a large, complex space, and not getting stuck on a local optimum. "Proposal mechanisms"
- MrBayes uses "metropolis coupled MCMC", or MCMCMC, to help explore the space. Here, there are (by default) 4 chains, 1 cold and 3 hot. The hot chains are allowed to vary more freely. If a chain happens to find a region with higher likelihood, then (again at a certain rate determined by R) the chain will become the new cold chain that gets sampled.
- Convergence: typically 2 analyses are run independently, and their convergence is measured. The runs will initially be far apart, and approach each other. This is the "burn-in" period. As a rule of thumb, you want convergence of the average standard deviation of split frequencies to get below 0.01.
- You can have convergence problems especially when searching a really big space, e.g. >150 taxa, or when your data just doesn't have enough signal to resolve the phylogeny.