IB200, Spring 2016 *University of California, Berkeley*

Lecture 12:

Bayesian phylogenetics and Markov chain Monte Carlo

Will Freyman

1 Basic Probability Theory

Probability is a quantitative measurement of the likelihood of an outcome of some random process. The probability of an event, like flipping a coin and getting heads is notated $P(heads)$. The probability of getting tails is then $1 - P(heads) = P(tails)$.

- The *joint probability* of event A and event B both occuring is written as $P(A, B)$. The joint probability of *mutually exclusive* events, like flipping a coin once and getting heads and tails, is 0.
- The probability of A occuring given that B has already occurred is written $P(A|B)$, which is read "the probability of A given B". This is a *conditional probability.*
- The *marginal probability* of A is $P(A)$, which is calculated by summing or integrating the joint probability over B. In other words, $P(A) = P(A, B) + P(A, \text{not } B)$.
- Joint, conditional, and marginal probabilities can be combined with the expression $P(A, B) = P(A)P(B|A) = P(B)P(A|B)$.
- The above expression can be rearranged into *Bayes' theorem*:

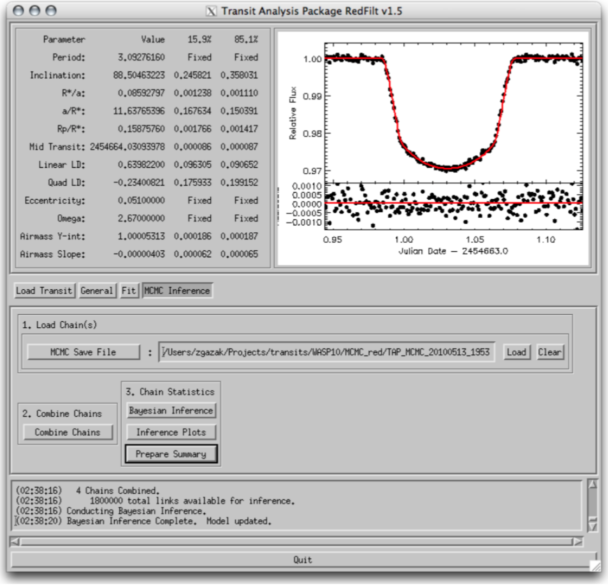
  $$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

  Bayes' theorem is a straightforward and uncontroversial way to calculate an inverse conditional probability.
- If $P(A|B) = P(A)$ and $P(B|A) = P(B)$ then A and B are *independent*. Then the joint probability is calculated $P(A, B) = P(A)P(B)$.

2 Interpretations of Probability

What exactly is a probability? Does it really exist? There are two major interpretations of probability:

- **Frequentists** believe that the probability of an event is its relative frequency over time. Probabilities are defined by a ratio from an infinite series of trials. Most stats taught to undergraduate biologists are frequentist, e.g. p-values, t-tests, null hyopthesis testing, etc. Maximum likelihood is a frequentist approach that maximizes $P(data|H)$ to get a **point estimate**.
- **Bayesians** measure the probability of an event as a degree of belief. Bayesians ask "How strongly can I believe in my hypothesis given this new data?" They update the prior probability of a hypothesis $P(H)$ with data to estimate the **posterior probability distribution** $P(H|data)$.



Parameter	Value	15.9%	85.1%
Period:	3.09276160	Fixed	Fixed
Inclination:	88.50463223	0.245821	0.358031
R*/a:	0.08592797	0.001238	0.001110
a/R*:	11.63765396	0.167634	0.150391
Rp/R*:	0.15875760	0.001766	0.001417
Mid Transit:	2454664.03093978	0.000086	0.000087
Linear LD:	0.63982200	0.096305	0.090652
Quad LD:	-0.23400821	0.175933	0.199152
Eccentricity:	0.05100000	Fixed	Fixed
Omega:	2.67000000	Fixed	Fixed
Airmass Y-int:	1.00005313	0.000186	0.000187
Airmass Slope:	-0.00000403	0.000062	0.000065

MCMC Save File : /Users/zgazak/Projects/transits/WASP10/MCMC_red/TAP_MCMC_20100513_1953

```
(02:38:16)   4 Chains Combined.
(02:38:16)      1800000 total links available for inference.
(02:38:16) Conducting Bayesian Inference.
(02:38:20) Bayesian Inference Complete.  Model updated.
```

Figure 1: Screenshot from the software Transit Analysis Package (TAP) [Gazak et al., 2012] that uses Bayesian inference and MCMC to discover exoplanets. Shown is the actual data for exoplanent WASP-10b published in Johnson et al. [2009].

3 Bayesian Inference in Science

We can rewrite Bayes' theorem to be:

$$P(H|D) = \frac{P(D|H)P(H)}{P(D)}$$

Here, H is a hypothesis (or model), and D is the observed data.

- The value $P(H|D)$ is the *posterior probability*, the probability of the hypothesis given the data. Intuitively, this is often what scientists want to know when testing hypotheses.

- $P(D|H)$ is the probability of the data given the hypothesis. This is called the *likelihood*. In maximum likelihood analysis this is the only term that is considered.

- $P(H)$ is the *prior probability* of the hypothesis, or the probability of the hypothesis before considering the data.

- The *marginal likelihood* $P(D)$ is the probability of the data marginalized over all hypotheses or models. This is sometimes a big, gnarly integral that is often impossible to analytically calculate.

- We have been treating H as a *discrete* hypothesis or parameter. If H is a *continuous* parameter then there are an infinite number of hypotheses because each specific parameter value is in some sense a separate hypothesis. Now we need to calculate a *probability density function* instead of a probability:

$$f(H|D) = \frac{f(D|H)f(H)}{\int f(D|H)f(H)dH}$$

Bayes' theorem essentially updates probabilities as new information is acquired. As new data is observed the prior probability is updated to produce a posterior probability.

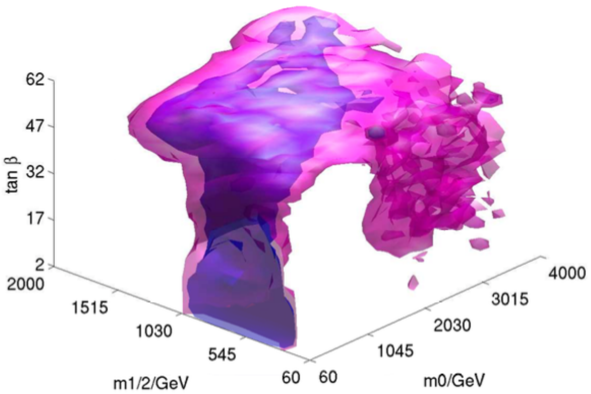


Figure 2: Bayesian inference and MCMC fit particle collider data to parameter rich supergravity models [Allanach et al., 2008].

Bayes' theorem has been around since at least 1763 [Bayes and Price, 1763], when the Reverend Thomas Bayes (1702-1761) first published it. However, applying Bayes' theorem to actual scientific inference was limited due to the difficulty of calculating the marginal likelihood, which could require integrating over hundreds or thousands of parameters. Consequently, statistician and biologist R.A. Fisher and others developed easier to compute frequentist approaches such as significance testing and maximum likelihood, which became the core of experimental science throughout the 20th century. However, within the last couple decades numerically approximating the marginal likelihood has been increasingly possible with algorithmic advances such as Markov chain Monte Carlo (MCMC) methods [Metropolis et al., 1953, Hastings, 1970] and growing computing power. Bayesian MCMC approaches were first developed by particle physicists modeling the dynamics of subatomic particles, and are now the standard tool used in astrophysics to discover exoplanents and in cosmology to model galaxy evolution (see Figures 1 and 2). They are a natural fit to any science – like evolutionary biology – that wishes to make inferences from large datasets about complex processes that may not be directly observable.

4 Bayesian Phylogenetics

In phylogenetics we are interested in the posterior probability of a tree given a character matrix. To calculate this we use Bayes' theorem as

$$f(\Psi, v, \theta | X) = \frac{f(X|\Psi, v, \theta) f(\Psi, v, \theta)}{f(X)}$$

where X is the character matrix, Ψ is a tree topology, v is a vector of branch lengths, and θ is a vector of character evolution model parameters. The likelihood $f(X|\Psi, v, \theta)$ is calculated using Felsenstein's pruning algorithm [Felsenstein, 1981], just like in maximum likelihood analyses and using the same models of character evolution. We will cover these models in lab this afternoon. $f(\Psi, v, \theta)$ is the prior probability of the tree, the branch lengths, and the character evolution parameters. We'll talk about picking priors below. Often we are primarily interested in the tree topology, so we can integrate out other parameters such as the character evolution model parameters by calculating

$$f(\Psi, v | X) = \frac{\int f(X|\Psi, v, \theta) f(\Psi, v, \theta) d\theta}{f(X)}$$

The marginal likelihood is calculated by summing over all tree topologies and integrating over all branch lengths and parameter values:

$$f(X) = \sum_{\Psi} \int_{\theta} \int_{v} f(X|\Psi, v, \theta) f(\Psi, v, \theta) d\theta dv$$

This high dimensionality integral is impossible to analytically compute, but we can still numerically approximate the posterior using Markov chain Monte Carlo algorithms (more on this later).

4.1 Posterior Distributions Represent Uncertainty

Bayesian statistics quantify uncertainty without the need for ad hoc procedures such as bootstrapping. In maximum likelihood and maximum parsimony analyses we are trying to estimate the single tree that maximizes the optimality criterion – this is a *point estimate*. In a Bayesian analysis we want to estimate the full posterior probability *distribution* of trees, and therefore retain all the information about phylogenetic uncertainty. If we use the full distribution of trees in downstream comparative analyses instead of relying on a single point estimate we will avoid falsely reducing the variance due to inflated phylogenetic certainty.

Of course sometimes it is necessary to summarize the posterior distribution of trees into a single tree (like for a figure in a paper). There are multiple ways to summarize tree distributions:

- The *maximum a posteriori* (MAP) tree is the tree topology that has the greatest posterior probability, averaged over all branch lengths and substitution parameter values. Given uninformative, flat priors, the MAP tree will often resemble the maximum likelihood tree.

- The *maximum clade credibility* (MCC) tree is the tree with the maximum product of the posterior clade probabilities.

- *50% Majority rule consensus tree* is a tree constructed so that it contains all of the clades that occur in at least 50% of the trees in the posterior distribution. This is very commonly used in the literature, but isn't the best option since it may be a topology that was never actually sampled during the analysis (it may have a probability of essentially zero).

Depending on the parameters of interest one could summarize the posterior probability distribution in many other ways as well, for example the set of 95% credible trees. However, a Bayesian only considers these as summaries of the true result of the analysis – which is the full posterior probability distribution.

4.2 Priors Are Useful and Good

Probably the most common criticism of Bayesian approaches is that they rely on priors. For example, Grant and Kluge [2003], which we read earlier this semester, asks about Bayesian priors:

> Should the beliefs of Creationists also be factored into phylogenetic inference? Or those of small children? Or the mentally ill? Should dreams and visions provide a basis for priors?

Actually priors are awesome:

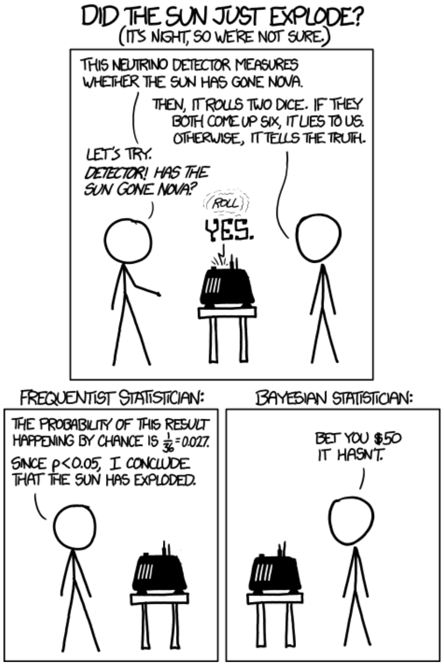

DID THE SUN JUST EXPLODE?
(IT'S NIGHT, SO WE'RE NOT SURE.)
THIS NEUTRINO DETECTOR MEASURES WHETHER THE SUN HAS GONE NOVA.
THEN, IT ROLLS TWO DICE. IF THEY BOTH COME UP SIX, IT LIES TO US. OTHERWISE, IT TELLS THE TRUTH.
LET'S TRY. DETECTOR! HAS THE SUN GONE NOVA?
ROLL
YES.
FREQUENTIST STATISTICIAN:
THE PROBABILITY OF THIS RESULT HAPPENING BY CHANCE IS $\frac{1}{36}$=0.027. SINCE p<0.05, I CONCLUDE THAT THE SUN HAS EXPLODED.
BAYESIAN STATISTICIAN:
BET YOU $50 IT HASN'T.

Figure 3: `http://xkcd.com/`

- Priors force us to make our modeling assumptions more explicit, unlike other methods such as parsimony in which the modeling assumptions are implicit.
- The influence of different priors on the resulting inference can be easily tested. This gives us a quantitative test of our modeling assumptions.
- Sometimes we have important prior information that we want to incorporate into our model – such as fossil ages (also see Figure 3). Priors give us the ability to include this information into our analyses.
- Using priors allow us to test the informativeness of our data. You can run the analysis under the prior only, without accounting for the data, and compare the results after running the analysis with the data. This assesses whether the data are informative for the parameters you are trying to infer.
- We can use a hyperprior (a prior on a prior) if we really have no idea what reasonable parameter value we should give our prior.

An *informative* prior has low variance, and is used to express specific information (e.g. a fossil aged 19-21 million years ago). A *noninformative* or *diffuse* prior is used when we have little or vague information. For example, in the absence of fossil data we might use a flat uniform prior distribution of 0-540 million years ago for the age of a clade.

4.3 Model Testing with Bayes Factors

We can compare two models and find the best-fitting model using Bayes factors. For any two models M_1 and M_2, the Bayes factor K is calculated as the posterior model odds divided by the prior model odds

$$K = \frac{P(M_1|D)}{P(M_2|D)} / \frac{P(M_1)}{P(M_2)}$$

A value of $K > 1$ means M_1 fits the data better than M_2. Like AIC, BIC, and likelihood-ratio tests, Bayes factors penalizes models for extra dimensions to avoid over fitting. Unlike AIC, BIC, and the likelihood-ratio test Bayes factors also takes into account the priors used in each model.

5 Markov chain Monte Carlo

Markov chain Monte Carlo (MCMC) algorithms represent a class of algorithms that approximate a multi-dimensional integral. They were developed to numerically approximate the posterior distribution in complex, parameter-rich hierarchical models. The are *Markov chains* in the sense that they are stochastic processes that transition from one state to another, and in which the probability distribution of the next state depends only on the current state and not on the sequence of events that preceded it. It is also a *Monte Carlo* method because it uses simulations based on random sampling.

5.1 MCMC algorithms

There are many different MCMC algorithms. Here I'll introduce just a few of the common ones used in phylogenetics.

5.1.1 Metropolis algorithm

The Metropolis algorithm was the first described MCMC algorithm [Metropolis et al., 1953]. The basic idea is this:

1. Initialize values for all parameters θ by drawing from the prior.
2. Propose new values for all parameters θ' by drawing from the proposal distribution.
3. Using the data X, calculate the acceptance ratio
$$R = \min\left[1, \frac{f(X|\theta')}{f(X|\theta)} \times \frac{f(\theta')}{f(\theta)}\right]$$
4. If $R \geq 1$ than we set $\theta = \theta'$, and go back to step 2. This means we always accept "uphill" steps.
5. Otherwise draw a random number $0 \leq \alpha \leq 1$. If $\alpha \leq R$ than we set $\theta = \theta'$, and go back to step 2. This means we *sometimes* accept "downhill" steps.
6. Otherwise discard θ' and go back to step 2.

As more values of θ are sampled, the full set of samples more closely approximates the posterior distribution of θ.

5.1.2 Metropolis-Hastings algorithm

The Metropolis-Hastings algorithm [Hastings, 1970] is the most commonly used MCMC algorithm. It extends the Metropolis algorithm by allowing for assymetric proposal distributions.

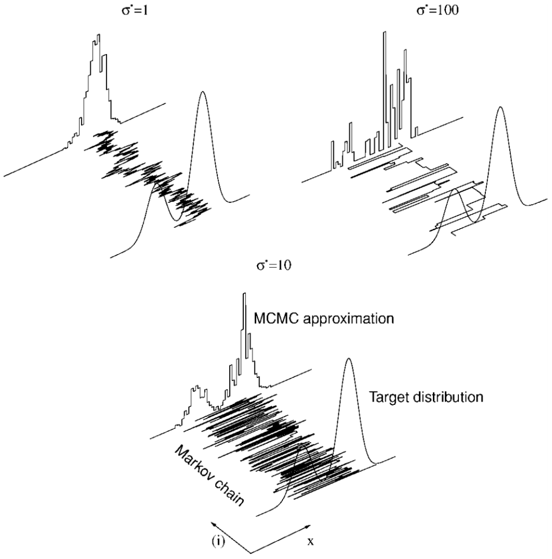


Figure 4: Approximations obtained using the Metropolis-Hastings algorithm with three Gaussian proposal distributions of different variances. From Andrieu et al. [2003].

5.1.3 Metropolis-coupled MCMC

Metropolis-coupled MCMC (MC^3) [Geyer, 1991] runs multiple parallel MCMC chains simultaneously. One of the chains is the *cold chain* which uses an acceptance ratio as described above. The other chains are *heated chains* in which the posterior probability is raised to a power β where $0 < \beta < 1$. β is referred to as the *temperature* of each chain. The heated chains accept new states more easily than the cold chain, allowing them to explore the peaks and valleys of parameter space more readily. Periodically the chains can swap states, so that the overall analysis more quickly converges on the target distribution.

5.1.4 Reversible-jump Markov chain Monte Carlo

Reversible-jump MCMC [rjMCMC; Green, 1995] samples across models of different dimensionality, drawing samples from models in proportion to their posterior probability and enabling Bayes factors for each model to be calculated. Moreover, we can decide not to rely on a single model but instead make model-averaged inferences by drawing estimates from all possible models weighted by their posterior probability.

5.2 Convergence and Mixing

How do you tell when an MCMC analysis has converged on the target distribution? How can you improve an MCMC anlysis so that it *mixes* (explores parameter space) more efficiently? This is often the trickiest part of Bayesian phylogenetic analysis, however this is really a limitation with our MCMC algorithms and computational resources, not Bayesian statistics itself. New MCMC algorithms are a highly active area of research, and many different diagnostics have been developed to assess convergence.

It is **extremely** important to ensure your MCMC analysis has converged. Visual inspection of trace plot is the quickest way to assess convergence (see Figure 4), but we'll discuss this and other convergence diagnotistics in lab next week.

References

Benjamin C Allanach, Matthew J Dolan, and Arne M Weber. Global fits of the large volume string scenario to wmap5 and other indirect constraints using markov chain monte carlo. *Journal of High Energy Physics*, 2008(08):105, 2008.

Christophe Andrieu, Nando De Freitas, Arnaud Doucet, and Michael I Jordan. An introduction to mcmc for machine learning. *Machine learning*, 50(1-2):5–43, 2003.

Mr. Bayes and Mr Price. An essay towards solving a problem in the doctrine of chances. by the late rev. mr. bayes, frs communicated by mr. price, in a letter to john canton, amfrs. *Philosophical Transactions (1683-1775)*, pages 370–418, 1763.

Joseph Felsenstein. Evolutionary trees from dna sequences: a maximum likelihood approach. *Journal of molecular evolution*, 17(6):368–376, 1981.

J Zachary Gazak, John A Johnson, John Tonry, Diana Dragomir, Jason Eastman, Andrew W Mann, and Eric Agol. Transit analysis package: An idl graphical user interface for exoplanet transit photometry. *Advances in Astronomy*, 2012, 2012.

Charles J Geyer. Markov chain monte carlo maximum likelihood. 1991.

Taran Grant and Arnold G Kluge. Data exploration in phylogenetic inference: scientific, heuristic, or neither. *Cladistics*, 19(5):379–418, 2003.

Peter J Green. Reversible jump markov chain monte carlo computation and bayesian model determination. *Biometrika*, 82(4):711–732, 1995.

W Keith Hastings. Monte carlo sampling methods using markov chains and their applications. *Biometrika*, 57(1):97–109, 1970.

John Asher Johnson, Joshua N Winn, Nicole E Cabrera, and Joshua A Carter. A smaller radius for the transiting exoplanet wasp-10bbased on observations obtained with the university of hawaii 2.2 m telescope operated by the institute for astronomy. *The Astrophysical Journal Letters*, 692(2):L100, 2009.

Nicholas Metropolis, Arianna W Rosenbluth, Marshall N Rosenbluth, Augusta H Teller, and Edward Teller. Equation of state calculations by fast computing machines. *The journal of chemical physics*, 21(6):1087–1092, 1953.