Bayesian methods

Ziheng Yang
Department of Biology
University College London

Plan

- Probability and principles of statistical inference
- Bayes's theorem & Bayesian statistics
- Bayesian computation
- Two applications
  - coalescent analysis of a DNA sample
  - phylogeny reconstruction

Probability: dual concepts

1. Frequency
   When I toss this coin 1000 times, the frequency of heads is about ½.
2. Degree of (rational or personal) belief
   The probability that it will rain tomorrow is ½.

Frequentist (classical) statistics

In Frequentist statistics, parameters are fixed, and we think of properties of estimation methods in repeated sampling, that is, when we imagine taking many data samples from the same process that generated our observed data.

It is not meaningful to talk about the probability that the parameter falls within a range, such as Prob($\theta > 0$), or the probability of a hypothesis, Prob(H_0).

Bayesian statistics

Probability measures degree of belief. Inference is conditional on the observed data. There is not much distinction between parameters and random variables.

Confidence interval (CI)

Suppose the data are a sample (x_1, x_2, …, x_n) from the normal distribution $N(\mu, \sigma^2)$, with unknown mean μ and variance σ^2. If n is large, the 95% confidence interval for μ is

$$(\bar{x} - 1.96s/\sqrt{n},\ \bar{x} + 1.96s/\sqrt{n})$$

It is incorrect to say that the CI includes the true mean with probability 95%.

A 75% confidence interval

Suppose we take two random draws (x_1 and x_2) from the following distribution to estimate θ ($-\infty < \theta < \infty$). The following procedure produces a 75% confidence interval (set).

$$\Pr(X = \theta - 1) = \Pr(X = \theta + 1) = \tfrac{1}{2}.$$

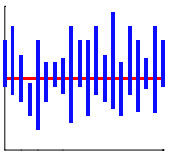
$$\hat{\theta} = \begin{cases} (x_1 + x_2)/2, & \text{if } x_1 \neq x_2, \\ x_1 - 1, & \text{if } x_1 = x_2. \end{cases}$$

Data outcomes
++: ✪
+−: ✪
−+: ✪
−−: ★

Before the experiment, the probability that the interval contains the true θ is 75%.
After the experiment, it is either 0 or 1.

Confidence interval (CI) vs. Bayesian credibility interval (CI)

The 95% confidence interval (θ_L, θ_U): Imagine that we fix θ and draw many data samples under this θ. In each sample, construct a 95% CI, which will vary among samples. Among those CIs, 95% of them cover the true θ. Sometimes the 95% CI from the observed data clearly does not include the true θ (that is, the probability that the CI includes θ is 0).



Given the data, the 95% Bayesian credibility interval (θ_L, θ_U) includes the true θ with probability 95%.

Bayes's theorem (inverse probability theorem)

Example (screening paradox). Suppose a person has tested positive in a clinical test. What is the probability that he has the infection?

***P*(positive | infection) = 0.99**
***P*(positive | no infection) = 0.02**
***P*(infection) = 0.001**
***P*(no infection) = 0.999**

Bayes's theorem

***P*(positive | infection) = 0.99**
***P*(positive | no infection) = 0.02**
***P*(infection) = 0.001**
***P*(no infection) = 0.999**

***P*(positive) = 0.001 × 0.99 + 0.999 × 0.02 = 0.02097**
***P*(infection | positive) = 0.001 × 0.99/0.02097 = 0.047**

Bayes's theorem

A: infection; $\bar{A}$: no infection

B: test-postive

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

Bayesian estimation of θ

$$f(\theta_i \mid x) = \frac{f(\theta_i) f(x \mid \theta_i)}{f(x)} = \frac{f(\theta_i) f(x \mid \theta_i)}{\sum_j f(\theta_j) f(x \mid \theta_j)}$$

$$f(\theta \mid x) = \frac{f(\theta) f(x \mid \theta)}{f(x)} = \frac{f(\theta) f(x \mid \theta)}{\int f(\theta) f(x \mid \theta)\, \mathrm{d}\theta}$$

The posterior is proportional to the prior times the likelihood. The posterior information is the sum of the prior information and the sample information.

$f(\theta)$: prior; $f(\theta|x)$: posterior; $f(x|\theta)$: likelihood; $f(x)$: normalizing constant

The use of Bayes's theorem when $f(\theta)$ does not have a frequency interpretation is controversial.

All controversies about Bayesian statistics are about the prior.
Bayesians claim that classical statistics is a fundamentally flawed theory with *ad hoc* fixes that often work, while Bayesian statistics is a fundamentally valid theory with some technical difficulties.

Bayesian credibility interval (CI)

The 95% credibility interval (θ_L, θ_U):

Let x_1, x_2, …, x_n be a sample from $N(\theta, 1)$. Assume a *non-informative* prior on θ. Then the 95% CI is

$$\bar{x} \pm 1.96/\sqrt{n}$$

Given the data, the Bayesian CI includes the true θ with probability 95%.

P value vs. posterior probability

Significance test: H_0: $\theta < 0$.

- *P* value is not the probability that H_0 is correct. It is the probability of observing data at least as extreme as the observed data if H_0 is correct.

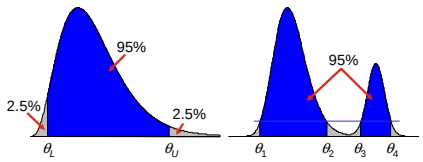
  P value = Pr(extreme data | H_0)

- Bayesian posterior probability for H_0 is the probability that H_0 is correct, given the data.

  Pr($\theta < 0$|data)

All Bayesian inference is based on the posterior.

- Mean, median, mode as point estimate
- 95% equal-tail credibility interval: (θ_L, θ_U)
- 95% highest posterior density (HPD) region (interval): (θ_1, θ_2), (θ_3, θ_4)



Example: Jukes-Cantor distance

data: *x* out of *n* sites are different.

$$L(\theta; x) = f(x;\theta) = \frac{n!}{x!(n-x)!}\, p^x (1-p)^{n-x}$$

$$p = \tfrac{3}{4}[1 - \exp(-\tfrac{4}{3}\theta)]$$

MLEs:

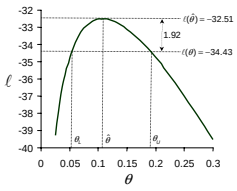
$$\hat{p} = \tfrac{x}{n}$$

$$\hat{\theta} = -\tfrac{3}{4}\log(1 - \tfrac{4}{3} \times \tfrac{x}{n})$$

Example: Jukes and Cantor distance

***x*=10 differences out of *n*=100 sites**

MLE and likelihood interval



The Bayesian solution

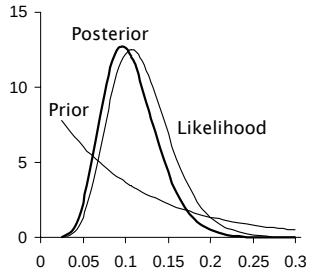
Suppose we use an exponential prior with mean $\mu = 0.1$.

$$f(\theta) = \tfrac{1}{\mu}\exp(-\tfrac{1}{\mu}\theta), \qquad 0 < \theta < \infty$$

$$f(\theta \mid x) = \frac{f(\theta) f(x \mid \theta)}{f(x)} = \frac{f(\theta) f(x \mid \theta)}{\int f(\theta) f(x \mid \theta)\, d\theta}$$

$$f(x \mid \theta) = \tfrac{n!}{x!(n-x)!}[\tfrac{3}{4} - \tfrac{3}{4}\exp(-\tfrac{4}{3}\theta)]^x \times [\tfrac{1}{4} + \tfrac{3}{4}\exp(-\tfrac{4}{3}\theta)]^{n-x}$$

The Bayesian solution: numerical integration



The prior $f(\theta)$

- It describes our previous knowledge about the parameter before data are considered (objective Bayesian)
- It reflects my personal belief about the parameter before the data are collected (subjective Bayesian)
- Difficulties in representing ignorance (*noninformative*, *vague*, *diffuse*, *reference* priors).
- Prior means your prejudice against mine as well as different inferences from the same data.

The difficulties of representing ignorance using uniform distributions

- Discrete case
  Prob(E occurs on weekend, not on weekday) = ½ or 2/7
- Continuous case (size of square)
  The side is $U(1, 2)$ meters
  The area is $U(1, 4)$ square meters

Ways for specifying priors

- Use of a physical model to describe uncertainties n parameters
- Previous data or knowledge under similar conditions
- Mathematical convenience (conjugate priors)
- vague (diffuse) prior
- Personal beliefs

Bayesian computation

- Difficulties in calculating high-dimensional integrals
- Markov chain Monte Carlo (MCMC)
- Application to molecular phylogenetics

Difficulty in calculating the integrals was a major reason that prevented the widespread use of Bayesian statistics.

Numerical integration (the curse of dimension)
Monte Carlo integration (& importance sampling)
Markov chain Monte Carlo

Monte Carlo integration

To calculate

$$I = E_f\{h(\theta)\} = \int h(\theta) f(\theta)\, \mathrm{d}\theta$$

where $f(\theta)$ is a density, draw independent samples $\theta_1, \theta_2, \ldots, \theta_N$ from $f(\theta)$. Then

$$\hat{I} = \frac{1}{N}\sum_{i=1}^{N} h(\theta_i)$$

$$\mathrm{var}\{\hat{I}\} = \frac{1}{N^2}\sum_{i=1}^{N}\left(h(\theta_i) - \hat{I}\right)^2$$

Monte Carlo integration: difficulties

- **We rarely know how to sample from the posterior.**
- **Sampling from the prior is inefficient.**

Markov chain Monte Carlo

Draw dependent samples $\theta_1, \theta_2, \ldots, \theta_N$ from $f(\theta|x)$ such that $\theta_1, \theta_2, \ldots, \theta_N$ form a time-homogeneous Markov chain. Then

$$\tilde{I} = \frac{1}{N}\sum_{i=1}^{N} h(\theta_i)$$

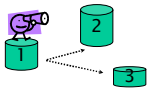
$$\mathrm{var}\{\tilde{I}\} = \mathrm{var}\{\hat{I}\} \times [1 + 2(\rho_1 + \rho_2 + \ldots)]$$

Metropolis algorithm for discrete parameter (Metropolis et al. 1953)

The algorithm generates a Markov chain with state space $\theta = 1, 2, 3$ and target density $\pi(\theta)$. (Suppose $\pi_1 = 0.3$, $\pi_2 = 0.5$, $\pi_3 = 0.2$, but we can calculate their ratios only.)

1. Set initial state: $\theta = 1$ (say).
2. Propose one of the two alternative states with equal probability ½. Let this be θ^*.
3. Accept or reject the proposal θ^*. If $\pi(\theta^*) > \pi(\theta)$, accept θ^*. Otherwise accept θ^* with probability $\pi(\theta^*)/\pi(\theta)$. If the proposal is accepted, set $\theta = \theta^*$. Otherwise set $\theta = \theta$. Print out θ.
4. Go to step 2.

1 2 1 1 3 2 2 2 1 2 2 3 2 2 2 1 ...



Features of the algorithm

- The proposal density is symmetrical: $q(\theta|\theta') = q(\theta'|\theta) = ½$.
- The sequence of states sampled over the iterations forms a Markov chain.
- The steady-state distribution of the chain is $\pi(\theta)$; that is, the time the boy spends on each box is proportional to the height of that box.
- The algorithm requires calculation of the ratio $\pi(\theta')/\pi(\theta)$, but not of $\pi(\theta)$.

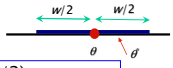
The ratio of the posterior is easier to calculate than the posterior itself

$$\pi(\theta) = f(\theta|x) = \frac{f(\theta) f(x|\theta)}{f(x)}$$

$$\frac{\pi(\theta^*)}{\pi(\theta)} = \frac{f(\theta^*) f(x|\theta^*)}{f(\theta) f(x|\theta)}$$

Metropolis algorithm (Metropolis et al. 1953) for a continuous parameter

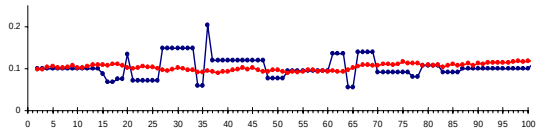
JC69 distance calculation, target density $\pi(\theta)=f(\theta|x)$.



1. Initialize: $n = 100$, $x = 10$, $w = 0.01$.
2. Set initial state: $\theta = 0.5$, say.
3. Propose a new state $\theta' \sim U(\theta-w/2, \theta+w/2)$. If $\theta^* < 0$, set $\theta^* = -\theta^*$.
4. Calculate the acceptance probability
$$\alpha = \min\left(1, \frac{\pi(\theta^*)}{\pi(\theta)}\right) = \min\left(1, \frac{f(\theta^*)f(x|\theta^*)}{f(\theta)f(x|\theta)}\right)$$
5. Accept or reject the proposal θ'. Draw $r \sim U(0,1)$. If $r<\alpha$ set $\theta = \theta'$. Otherwise set $\theta = \theta$. Print out θ.
6. Go to step 3.

.5 .495 .495 .490 .491 .487 .479 .479 .479 ...

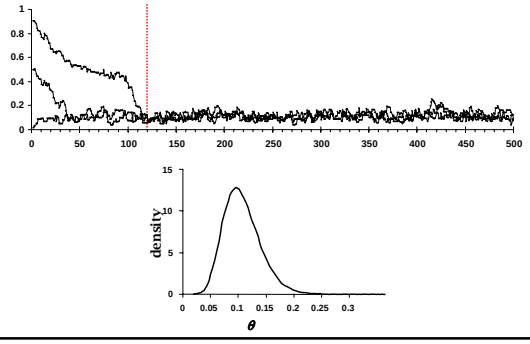
Neither large nor small windows are good.



w = 0.01, acceptance rate = 97%
w = 1, acceptance rate = 20%

Optimum acceptance rate is ~50% for 1-D proposal, decreasing to ~26% for multi-dimensional proposal.
Recommended values are 20-70% for 1-D and 15-40% for multi-D proposals.

Burn-in, histogram, density smoothing



Metropolis-Hastings algorithm (Hastings 1970)

The proposal (jump) density $q(\theta'|\theta)$ may be asymmetrical.
The acceptance probability is then

$$\alpha = \min\left(1, \frac{\pi(\theta^*)}{\pi(\theta)} \times \frac{q(\theta|\theta^*)}{q(\theta^*|\theta)}\right)$$
$$= \min\left(1, \frac{f(\theta^*)}{f(\theta)} \times \frac{f(x|\theta^*)}{f(x|\theta)} \times \frac{q(\theta|\theta^*)}{q(\theta^*|\theta)}\right)$$
$$= \min(1, \text{ prior ratio} \times \text{likelihood ratio} \times \text{proposal ratio})$$

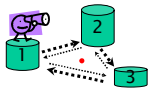
$$\pi(\theta) = f(\theta)f(x|\theta)/f(x)$$

Proposal ratio (Hastings ratio)

Suppose the robot proposes a left move with probability 2/3 and a right move with probability 1/3. By accepting left moves less often than right moves through the proposal ratio, the chain converges to the correct target distribution.

Example: $\theta = 1$, $\theta' = 2$.
$q(\theta|\theta') = 1/3$, $q(\theta'|\theta) = 2/3$,
$q(\theta|\theta')/q(\theta'|\theta) = ½$.

$$\alpha = \min\left(1, \frac{\pi(\theta^*)}{\pi(\theta)} \times \frac{q(\theta|\theta^*)}{q(\theta^*|\theta)}\right)$$



Proposals

The proposal (jump) density $q(\theta'|\theta)$ should specify a recurrent aperiodic Markov chain. It should be possible to reach any other state from any state, and the chain should not have a period.

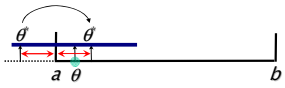
The proposal density can be entirely unrelated to the target density, so the same proposals can be used in different MCMC algorithms. The proposal greatly affects the convergence and mixing properties of the Markov chain.

Sliding window with reflection

$\theta^* \sim U(\theta - w/2, \theta + w/2)$

Suppose θ is defined in the interval (a, b). If the proposed value θ^* is outside the range, the excess is reflected back into the interval. This is a symmetrical proposal and the proposal ratio is 1.

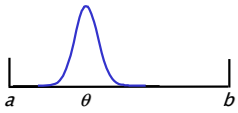
If $\theta^* < a$, reset θ^* to $a + (a - \theta^*) = 2a - \theta^*$.
If $\theta^* > b$, reset θ^* to $b - (\theta^* - b) = 2b - \theta^*$.



Sliding window with normal proposal

$\theta^* \sim N(\theta, \sigma^2)$

σ controls the step size.
If the proposal is outside the range, reflect as in the case of the uniform proposal.



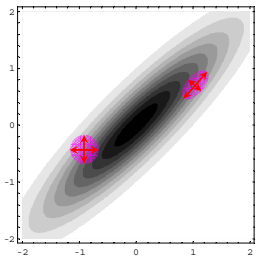
Correlation between parameters

Inefficient proposals

- one component at a time
- both components but ignoring the correlation

Efficient proposals

- reparametrize the model
- multi-dimensional proposal to account for correlation



Single-component M-H algorithm

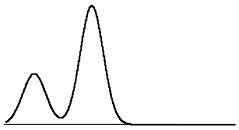
Partition multiple parameters into blocks: $\theta_1, \theta_2, \ldots, \theta_m$, each of which can be multi-dimensional.

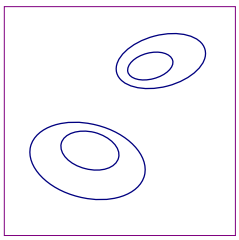
Propose changes to each block in turn, or update blocks with fixed probabilities.

It is more efficient to group highly-correlated parameters in one block and update them simultaneously.

Multiple local peaks

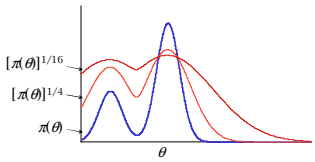
Difficult to cross valleys.





Metropolis-coupled Markov chain Monte Carlo (MCMCMC or MC^3)

MCMCMC runs several chains simultaneously, with one cold chain approaching the target while the other hot chains to help with the move.



Monitoring and diagnosing MCMC algorithms

Slow convergence and poor mixing are the two major problems.

- Use time series (trace) plot of variables. Check for convergence in "all" variables.
- Acceptance rate should be neither too high nor too low.
- Without data, the posterior should equal the prior.
- Use simulation to confirm target distribution.
- Should we run multiple long chains or one extremely long chain?

Excitements about MCMC?

MCMC has revolutionized Bayesian statistics in the past two decades. It offers exciting opportunities for implementing sophisticated and realistic models for analysis of genetic data.

Nevertheless, MCMC algorithms are difficult to code and validate. The problem is exacerbated by the use of parameter-rich models which are hardly identifiable.

MCMC algorithms are part science part art!

	Likelihood optimization	Bayesian MCMC
Likelihood	always goes up	no direction
Gradient	goes to 0	no direction
Convergence	to a point (MLEs)	to a distribution
Ways to make mistakes	many	more
Finding bugs	difficult	more difficult

Likelihood vs. Bayesian

	Likelihood (frequentist)	Bayesian
Invariant to parameterizations?	MLEs are	prior is not
Prior	No, thanks.	Yes, please.
Nuisance parameters	problematic	straightforward
Inference	conditional on parameters, indirect Frequentist interpretation	inference conditional on data, straightforward interpretation

Application 1: The neutral coalescent

Classic population genetics theory studies the change of gene frequencies over generations, influenced by random sampling (genetic drift), natural selection, etc.

Fisher R. 1930. *The Genetic Theory of Natural Selection*. Clarendon Press, Oxford.
Haldane JBS. 1932. *The Causes of Evolution*. Longmans Green & Co., London.
Wright S. 1931. Evolution in Mendelian populations. *Genetics* 16:97-159.

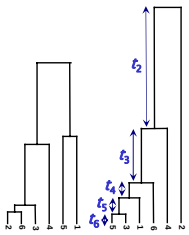
Modern work (*a*) is dominated by data, (*b*) uses the coalescent model, which "runs the time machine backward", and (*c*) is often computation-intensive (MCMC).

Kingman JFC. 1982. On the genealogy of large populations. *J. Appl. Prob.* 19A:27-43.
Kingman JFC. 1982. The coalescent. *Stochastic Process Appl.* 13:235-248.

Hein J, Schieriup MH, Wiuf C. 2005. *Gene Genealogies, Variation and Evolution: A Primer in Coalescent Theory*. Oxford University Press, Oxford.

Wakeley J. 2007. *Coalescent Theory: An Introduction*. Roberts & Company.

The coalescent model ($\theta = 4N\mu$)



Measure time in N generations and look backward in time. Then neutral mutations accumulate at rate $\theta/2$ while coalescent events occur at rate 1 for each pair of lineages.
Each genealogy (G) has equal probability. The waiting times (t_j) until the next coalescence have independent exponential distributions:

$$f(t_j) = \frac{j(j-1)}{2} \exp\left(-\frac{j(j-1)}{2} t_j\right)$$

Estimation of $\theta = 4N\mu$ from a population sample at a neutral locus

$$f(\theta \mid X) \propto \sum_i \int f(\theta) f(G_i) f(\mathbf{t}_i \mid \theta, G_i) f(X \mid \theta, G_i, \mathbf{t}_i)\, d\mathbf{t}_i$$

Random variables integrated out in the model:

- genealogy (tree topology) G_i
- $s - 1$ coalescent times $\mathbf{t}_i$ on each G_i

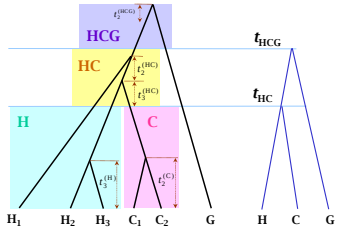
Sketch of an MCMC algorithm

- Start with a random tree G, with random coalescent times $\mathbf{t}$, and random θ.
- Each iteration consists of the following:
  - Propose a change to the tree, by rearranging nodes, which may change times $\mathbf{t}$ as well.
  - Propose a change to the times $\mathbf{t}$.
  - Propose a change to parameter θ.
- Every k iterations, sample the chain: save θ as well as G and $\mathbf{t}$ to disk.
- After many iterations, summarize the results (mean, median of θ, and other features of the posterior.

Population sizes and species divergence times

Parameters:

- **Speciation times:** t_{HC}, t_{HCG}
- **Population sizes:** θ_H, θ_C, θ_{HC}, θ_{HCG}



Yang (2002. Genetics 162:1811-1823)
Rannala & Yang (2003. Genetics 164:1645-1656)

Estimation of $\theta = 4N\mu$ from a population sample at a neutral locus

Kuhner, Yamato & Felsenstein (1995. Genetics 140:1421-1430) uses an MCMC algorithm to calculate the likelihood for given θ under a finite-site model, using θ_0 as a driving value. (coalesce, migrate, recombine → lamarck)

Stephens & Donnelly (2000 J. R. Statist. Soc. B. 62:605-655) discussed problems with the idea of using a driving value θ_0 to derive likelihood at other values of θ.

Estimation of $\theta = 4N\mu$ at a neutral locus from a sample of DNA sequences

Griffiths & Tavare assume the infinite-site model of mutation, and an importance-sampling algorithm to calculate the likelihood.

Felsenstein, Kuhner, Yamato & Beerli (1999. IMS Lect. Notes Monogr. Ser. 33:163-185)

MCMC algorithms for closely related species/populations

Wilson, Weal & Balding (2003. *J. R. Statist. Soc. A* 166:155-201) deals with micro-satellite data. (Batwing)

Nielsen (2000. *Genetics* 154:931-942) models the divergence between two species followed by gene flow. The algorithm works on sequence data and a tree of 2 species. Hey & Nielsen (2004 *Genetics* 167: 747-760) extends this to multiple loci. (IM)

Beerli & Felsenstein (2001. Proc. Natl. Acad. Sci. U.S.A. 98:4563-4568) and Bahlo & Griffiths (2000. *Theor. Popul. Biol.* 57:79-95) assume an equilibrium model of migration among populations. (migrate)

Application 2: Bayesian phylogenetics

Edwards (1970. *J. R. Stat. Soc. B.* 32:155-174) discussed the conditional distribution of *labelled histories* for human populations given the data of gene frequencies. Edwards & Cavalli-Sforza used a Brownian motion to model the drift of transformed gene frequencies over time and used the Yule process to specify the distribution of labelled histories and the divergence times.

Bayesian phylogenetics: brief history

Three groups introduced the Bayesian methodology to estimation of molecular phylogenies:

Rannala & Yang (1996. J. Mol. Evol. 43:304-311)
Yang & Rannala (1997. Mol. Biol. Evol 14:717-724)
Mau & Newton (1997. J. Comput. Graph. Stat. 6:122-131)
Li, Pearl & Doss (2000. J. Amer. Stat. Assoc. 95:493-508)

Molecular clock is assumed.
Prior on tree is uniform or from the birth-death process with species sampling.

Bayesian phylogenetics: brief history

BAMBE
(Larget & Simon. 1999. Mol. Biol. Evol. 16:750-759)

MrBayes
(Huelsenbeck & Ronquist. 2001. Bioinformatics 17:754-755;
Ronquist & Huelsenbeck. 2003. Bioinformatics 19:1572-1574)

Molecular clock relaxed.

More efficient proposal algorithms are implemented.

More models are implemented.

Bayesian phylogenetics

$$P(\tau_i \mid X) \propto f(\tau_i) f(X \mid \tau_i)$$

$$P(\tau_i \mid X) = \frac{\iint f(\theta) f(\tau_i) f(\mathbf{b}_i \mid \theta, \tau_i) f(X \mid \theta, \tau_i, \mathbf{b}_i) \mathrm{d}\mathbf{b}_i \mathrm{d}\theta}{\sum_j \iint f(\theta) f(\tau_j) f(\mathbf{b}_j \mid \theta, \tau_j) f(X \mid \theta, \tau_j, \mathbf{b}_j) \mathrm{d}\mathbf{b}_j \mathrm{d}\theta}$$

Parameters that need priors:

- tree topology τ_i: uniform
- branch lengths $\mathbf{b}_i$: $U(0,10)$ or exponential
- parameters in the substitution model θ

Sketch of an MCMC algorithm

- Start with a random tree τ, with random branch lengths $\mathbf{b}$, and random substitution parameters θ.
- In each iteration do the following:
  - Propose a change to the tree, by using tree rearrangement algorithms (such as nearest neighbour interchange or subtree pruning and regrafting). The step may change $\mathbf{b}$ as well.
  - Propose changes to branch lengths $\mathbf{b}$.
  - Propose changes to parameters θ.
- Every k iterations, sample the chain: save τ, $\mathbf{b}$, θ to disk.
- At the end of the run, summarize the results.

Bayesian phylogenetics: summaries

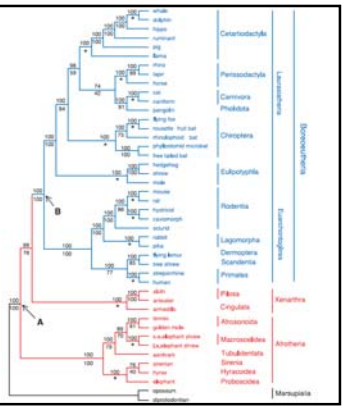
- MAP tree: tree topology with the maximum posterior probability.
- 95% credibility set of trees includes trees with the highest posterior probabilities until the total probability exceeds 95%.
- Posterior clade probability: proportion of sampled trees that contain the clade, shown on the majority-rule consensus tree

High posterior probabilities

from Murphy et al. (2001. Science 294:2348–2351)

16.4K bp

Posterior
bootstrap ML



Posterior probabilities for trees and clades appear too high and in general are not due to convergence problems with the MCMC.

If the prior and likelihood model are both correct, the posterior probabilities are indeed the probabilities that the tree or clade is correct, as theory predicts.

The posterior probabilities appear sensitive to model misspecifications, and to prior about (internal) branch lengths, and vague (diffuse) priors lead to extreme probabilities.

Bayesian model selection with vague priors on parameters is a difficult and controversial area.

Further reading

Yang, Z. 2006 *Computational Molecular Evolution*, OUP. Chapter 5

DeGroot, M. H., and M. J. Schervish. 2002. Probability and Statistics. Addison-Wesley, Boston, USA.

Leonard, T., and J. S. J. Hsu. 1999. Bayesian Methods. Cambridge University Press, Cambridge.

Gilks, W. R., S. Richardson, and D. J. Spielgelhalter. 1996. Markov Chain Monte Carlo in Practice. Chapman and Hall, London.