Bayesian phylogenetics: an introduction

Marc A. Suchard
msuchard@ucla.edu

UCLA
Who is this man?

How sure are you?
The one ‘true’ tree?

- Methods we’ve learned so far try to find a **single** tree that best describes the data.
- However, they do not search everywhere, and it is difficult to find the “best” tree.
- Many (gazillions of) trees may be almost as good.
Bayesian phylogenetics: general principle

- Using Bayesian principles, we will search for and average over sets of plausible trees (weighted by their probability) instead of a single “best” tree.

- In this method, the “space” that you search is limited by prior information and the data.

- The posterior distribution of trees naturally translates into probability statements (and uncertainty) on aspects of direct scientific interest:
  - When did an evolutionary event happen?
  - Are a subset of sequences more closely related?

- The cost: we must formalize our prior beliefs.
Linda is 31 years old, single, outspoken, and very bright. She majored in philosophy. As a student, she was deeply concerned with issues of discrimination and social justice, and also participated in antinuclear demonstrations.

Which is more probable?

1. Linda is a bank teller.
2. Linda is a bank teller and is active in the feminist movement.
Conditional probability: intuition

Conjunction (joint probably) fallacy:
- Tversky and Kahneman (1982)
- 85% of undergraduates chose the joint

I am particularly fond of this example [the Linda problem] because I know that the [conjoint] statement is least probable, yet a little homunculus in my head continues to jump up and down, shouting at me – “but she can’t just be a bank teller; read the description.”

- Stephen J. Gould
Conditional probability: intuition

- Arbitrary events $A$ (teller) and $B$ (activist) from sample space $U$
Bayes theorem

Definition of conditional probability in words:

\[ \text{probability}(A \text{ and } B) = \text{probability}(A \text{ given } B) \times \text{probability}(B) \]

In usual mathematical symbols:

\[ p(A|B)p(B) = p(A, B) = p(B|A)p(A) \]

With a slight re-arrangement:

\[ p(A|B) = \frac{p(B|A)p(A)}{p(B)} \]

“Just" a restatement of conditional probability
Bayes theorem

Integration (averaging) yields a marginal probability:

$$p(A) = \int p(A, B) dB = \int p(A|B)p(B) dB$$

over all possible values of $B$

- probability(teller) = probability(teller and activist) + probability(teller and not activist)
Conditional probability: pop quiz

What do you know about Thomas Bayes? Bayes theorem?

Some discussion points:
- Favorite game? Best buddies?
Bayes theorem for statistical inference

- Unknown quantity $\theta$ (model parameters, scientific hypotheses)
- Prior $p(\theta)$ beliefs before observed data $Y$ become available
- Conditional probability $p(Y|\theta)$ of the data given fixed $\theta$ – also called the likelihood of $Y$
- Posterior $p(\theta|Y)$ beliefs:

\[
p(\theta|Y) = \frac{p(Y|\theta)p(\theta)}{p(Y)}
\]

- $p(\theta)$ and $p(Y|\theta)$ – easy
- $p(Y) = \int p(Y|\theta)p(\theta)d\theta$ – hard
Bayesian phylogenetic inference
Bayesian phylogenetic inference

- Posterior:
  \[ p(\theta | Y) = \frac{p(Y | \theta)p(\theta)}{p(Y)} \]

- **Trouble**: \( p(Y) \) is not computable – sum over all possible trees

  For \( N \) taxa: there are \( G(N) = (2N - 3) \times (2N - 5) \times \cdots \times 1 \)

- \( \theta = \) (tree, substitution process)

  \( p(Y | \theta) \) - continuous-time Markov chain process that gives rise to sequences at tips of tree

  E.g., \( G(21) > 3 \times 10^{23} \)
• Strongest assumption: most parameters are separable, e.g. the tree is independent of the substitution process
• Weaker assumption: tree $\sim$ Coalescent process
• Weaker assumption: functional form on substitution parameters

- Normal distribution
- Exponential distribution
- Uniform distribution
- Gamma distribution
- Witch's Hat distribution

- Specialized priors as well
- If worried: check sensitivity
Posterior inference

Numerical (Monte Carlo) integration as a solution:
Markov chain Monte Carlo

- Metropolis et al (1953) and Hastings (1970) proposed a stochastic integration algorithm that can explore vast parameter spaces.
- Algorithm generates a Markov chain that visits parameter values (e.g., a specific tree) with frequency equal to their posterior density / probability.
- Markov chain: random walk where the next step only depends on the current parameter state.
Metropolis-Hastings Algorithm

- Each step in the Markov chain starts at its current state $\theta$ and proposes a new state $\theta^*$ from an arbitrary proposal distribution $q(\cdot|\theta)$ (transition kernel).
- $\theta^*$ becomes the new state of the chain with probability:

$$R = \min \left( 1, \frac{p(\theta^*|Y) \times q(\theta|\theta^*)}{p(\theta|Y) \times q(\theta^*|\theta)} \right)$$

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

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

- Otherwise, $\theta$ remains the state of the chain.
Posterior sampling

We repeat the process of proposing a new state, calculating the acceptance probability and either accepting or rejecting the proposed move millions of times.

Although correlated, the Markov chain samples are valid draws from the posterior; however . . .

Initial sampling (burn-in) is often discarded due to correlation with chain’s starting point ($\neq$ posterior)
Often we propose changes to only a small number of dimensions in $\theta$ at a time (Metropolis-within-Gibbs).

In phylogenetics, mixing (correlation) in continuous dimensions is much better (smaller) than for the tree.

So, dominant approach has been keep-it-simple-stupid – alternatives exist and may become necessary:
- Gibbs sampler; slice sampler; Hamiltonian MC.
Tree Transition Kernels

narrow exchange

wide exchange
Posterior Summaries

For continuous $\theta$, consider:

- posterior mean or median $\approx$ MCMC sample average or median
- quantitative measures of uncertainty, e.g. high posterior density interval

For trees, consider:

- scientifically interesting posterior probability statement, e.g. the probability of monophyly $\approx$ MCMC sample proportion under which hypothesis is true
Posterior Probabilities
Summarizing Trees

Tree 1

Tree 2

Tree 3

Strict consensus tree

Majority-rule consensus tree
MCMC Diagnostics: within a single chain

- Visually inspect MCMC output traces.
- Measure autocorrelation within a chain: the effective sample size (ESS).
MCMC Diagnostics: across multiple chains

- Visually inspect MCMC output traces

Comparing different chains → variance among and between chains
Improving Mixing

(Only if convergence diagnostics suggest a problem)

- Run the chain longer
- Use a more parsimonious model (uninformative data)
- Change tuning parameters of transition kernels to bring acceptance rates to 10% to 70%
- Use different transition kernels (consult an expert)
Improving Mixing

[SISMID University of Washington Bayesian phylogenetics]
Why Bother being Bayesian?

In practice, we have almost no prior knowledge for the model parameters. So, why bother with Bayesian inference?

- Analysis provides directly interpretable probability statements given the observed data.
- MCMC is a stochastic algorithm that (in theory) avoids getting trapped in local sub-optimal solutions.
- Search space under Coalescent prior is astronomically “smaller.”
- By numerically integrating over all possible trees, we obtain marginal probability statements on hypotheses of scientific interest, e.g. specific branching events or population dynamics, avoiding bias.