Introduction to BEAST

Louise Moncla

Genomics of Disease in Wildlife Workshop

June 12, 2023

Goals for this lecture and lab:

1. What is difference between BEAST and Nextstrain, and how do I decide which one to use?

2. What BEAST is doing, what is an MCMC chain, and how do I tell if it is working?

3. How do I make decisions about picking models and priors?

4. How, functionally, do I use this piece of software and interpret results?

What is BEAST and why would I use it?

BEAST

Bayesian Evolutionary Analysis Sampling Trees

* The underlying probability is based on Bayes Theorem
* Includes information about priors
* Results are distributions, rather than point estimates
* Slow

* Meant to be flexible
* An overwhelming array of evolutionary and epidemiologic models

* Everything is inferred with Markov chain Monte Carlo (MCMC)
* Parameters are sampled and evaluated probabilistically

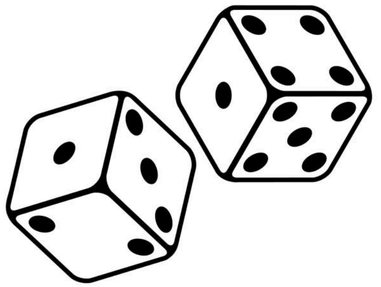
Conditional probabilities

A **conditional probability** is a measure of the probability of an event occurring, given that another event has already occurred.

Conditional probabilities

NOT a conditional probability

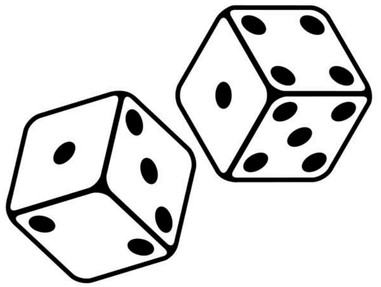
Pr (the number you rolled is a 6) = **1/6**



Conditional probabilities

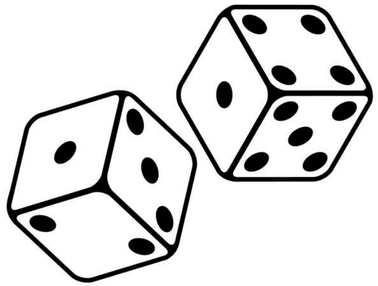
NOT a conditional probability

Pr (the number you rolled is a 6) = **1/6**



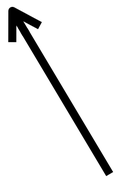
IS a conditional probability

Pr (the number you rolled is a 6, given that you know the number was even) = **1/3**



Bayes Theorem allows us to calculate conditional probabilities

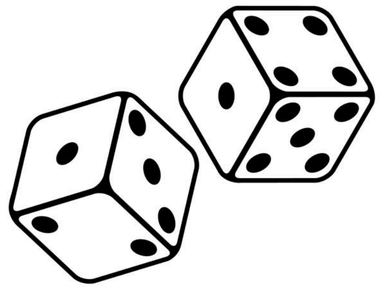
$$\Pr(A \mid B) = \frac{\Pr(B \mid A) \times \Pr(A)}{\Pr(B)}$$



This is pronounced: “the probability of A, given B”

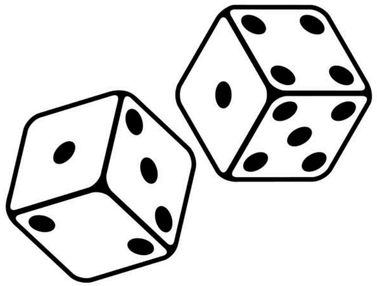
Bayes Theorem allows us to calculate conditional probabilities

$$\Pr(\text{roll 6} \mid \text{roll was even}) = \frac{\Pr(\text{roll was even} \mid \text{roll 6}) \times \Pr(\text{roll 6})}{\Pr(\text{roll was even})}$$



Bayes Theorem allows us to calculate conditional probabilities

$$\Pr(\text{roll 6} \mid \text{roll was even}) = \frac{\Pr(\text{roll was even} \mid \text{roll 6}) \times \Pr(\text{roll 6})}{\Pr(\text{roll was even})}$$



$$= \frac{1 \times 1/6}{1/2} = \mathbf{1/3}$$

Bayesian phylogenetics

Bayesian phylogenetics uses Bayes theorem to evaluate probabilities of tree topologies, given data and substitution models.

Bayes Theorem for phylogenetics

$$\Pr(\text{Parameters} \mid \text{Data}) = \frac{\Pr(\text{Data} \mid \text{Parameters}) \times \Pr(\text{Parameters})}{\Pr(\text{Data})}$$

*** Parameters = tree topology, branch lengths**

*** Data = multi-sequence alignment**

Bayes Theorem for phylogenetics

Likelihood of the alignment, given the parameters (tree topology, branch lengths)

Prior probability of the parameters (topology, branch lengths)

$$\Pr(\text{Parameters} \mid \text{Data}) = \frac{\Pr(\text{Data} \mid \text{Parameters}) \times \Pr(\text{Parameters})}{\Pr(\text{Data})}$$

Marginal probability of the alignment

*** Parameters = tree topology, branch lengths**

*** Data = multi-sequence alignment**

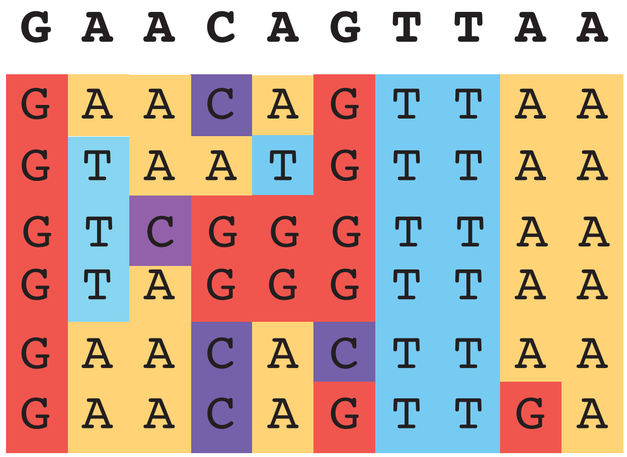
Markov chain Monte Carlo

- A Markov chain or Markov process is a stochastic model describing a sequence of possible events in which the probability of each event depends only on the state attained in the previous event.

- Monte Carlo methods are a broad class of computational algorithms that rely on repeated random sampling to obtain numerical results.

Markov chain Monte Carlo in BEAST

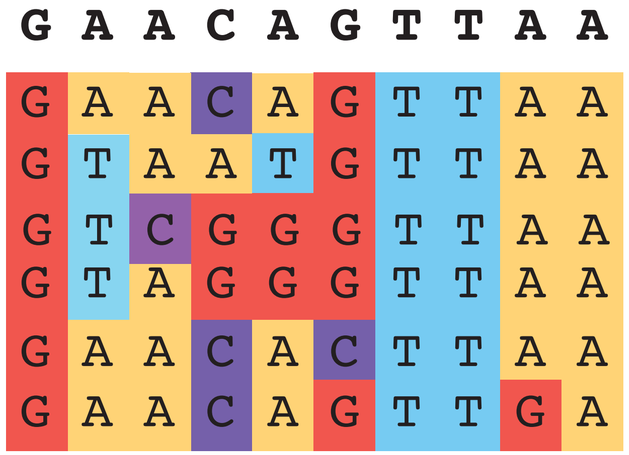
Input



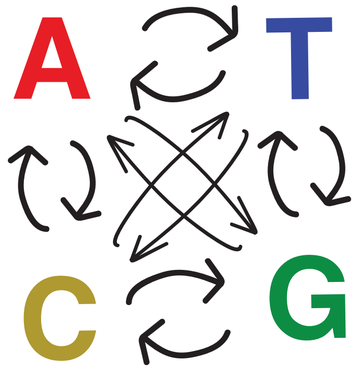
Alignment
(protein,
nucleotide, dates)

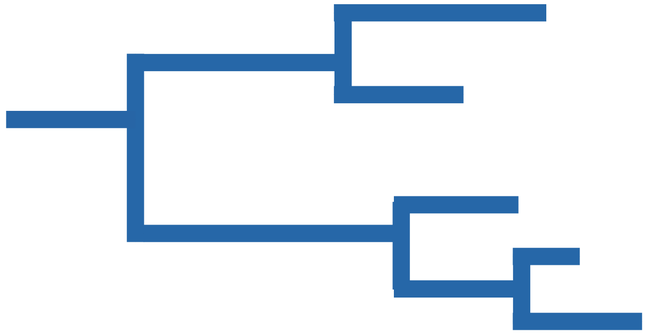
Markov chain Monte Carlo in BEAST

Input



What we're estimating





Alignment
(protein, nucleotide, dates)

Substitution model
(JC, HKY, GTR)

Tree model
(Coalescent, constant size, skyline, structured, etc…)

Markov chain Monte Carlo in BEAST

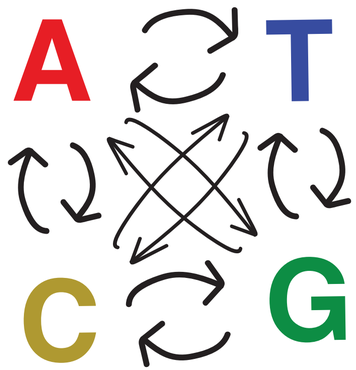
Input

G A A C A G T T A A

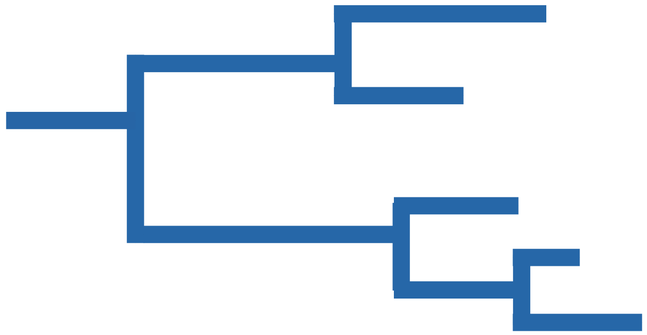
G	A	A	C	A	G	T	T	A	A
G	T	A	A	T	G	T	T	A	A
G	T	C	G	G	G	T	T	A	A
G	T	A	G	G	G	T	T	A	A
G	A	A	C	A	C	T	T	A	A
G	A	A	C	A	G	T	T	G	A

Alignment
(protein, nucleotide, dates)

What we're estimating



Substitution model
(JC, HKY, GTR)

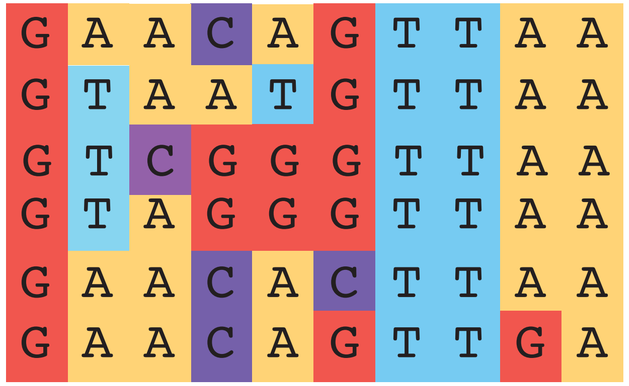


Tree model
(Coalescent, constant size, skyline, structured, etc…)

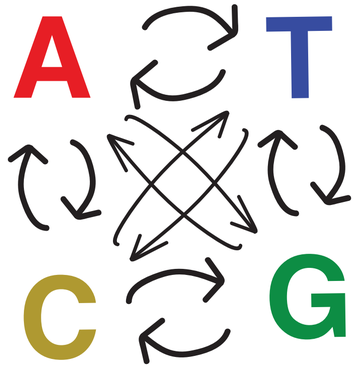
Markov chain Monte Carlo in BEAST

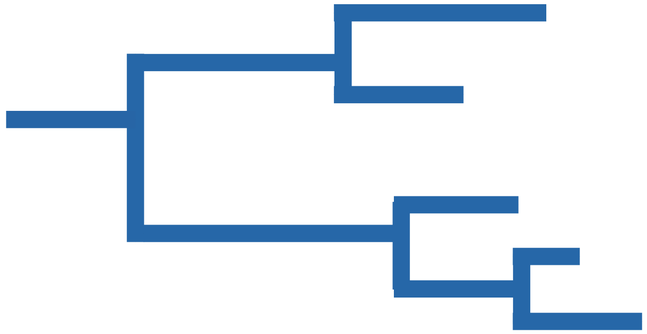
Input

G A A C A G T T A A



What we're estimating





Alignment

(nucleotide)

Substitution model

(HKY)

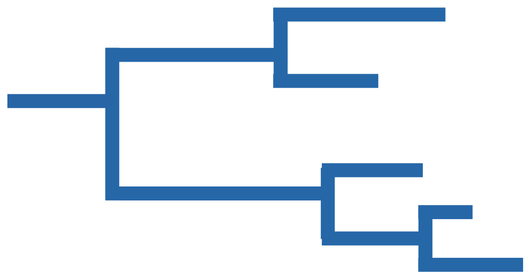
6 parameters: Transition rate, transversion rate, base frequencies for each base

Tree model

(Coalescent, constant size)

3 parameters: Ne, topology, branch lengths

MCMC step 1a: pick random values for each parameter



C/T ↷ A/G

C/T ↶ A/G

Freq(A)

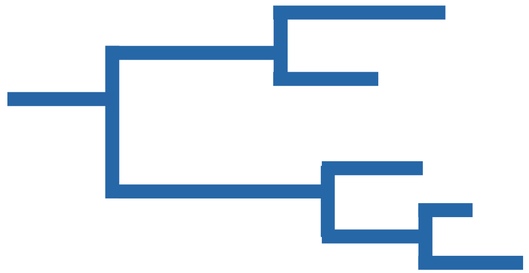
Freq(G)

Freq(C)

Freq(T)

Population size

MCMC step 1a: pick random values for each parameter



0.60

0.35

0.25

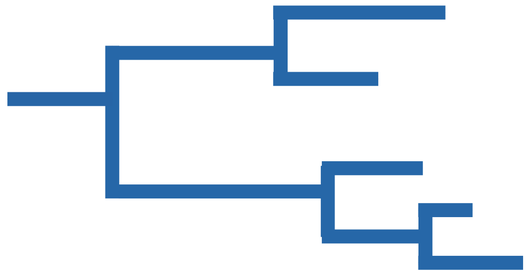
0.27

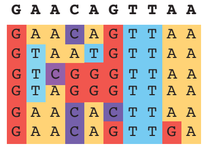
0.20

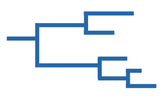
0.20

1000

MCMC step 1a: pick random values for each parameter



Pr(

|  , C/T ⇝ A/G C/T ⇜ A/G , Freq(A) Freq(C) Freq(G) Freq(T) , Population size)

0.60

0.35

0.25

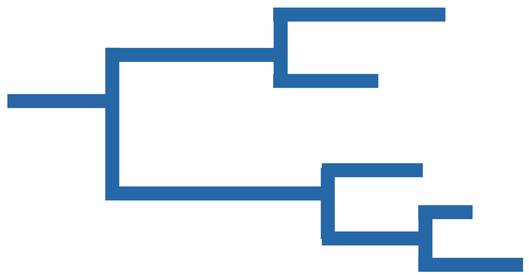
0.27

0.20

0.20

1000

MCMC step 1b: evaluate the probability of the alignment, given the sampled parameters



0.60

0.27

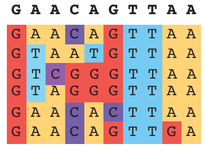
0.35

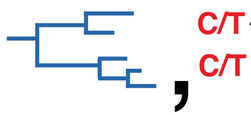
0.20

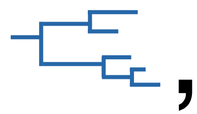
0.25

0.20

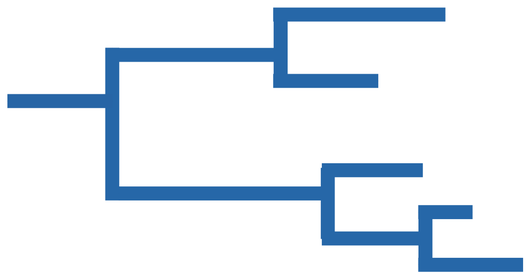
1000

Pr(

|  , C/T ⇝ A/G, C/T ⇜ A/G , Freq(A) Freq(C) Freq(G) Freq(T) , Population size)

* priorPr( , C/T ⇝ A/G, C/T ⇜ A/G , Freq(A) Freq(C) Freq(G) Freq(T) , Population size)

MCMC step 1b: evaluate the probability of the alignment, given the sampled parameters



0.60

0.27

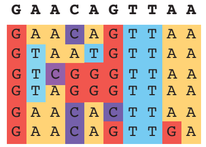
0.35

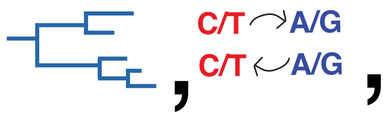
0.20

0.25

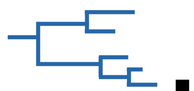
0.20

1000

Pr(

| 

)

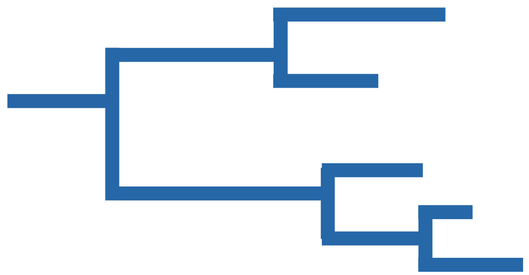
* priorPr(

)

= 10

* this represents the numerator of the equation on slide 12

MCMC step 2a: randomly sample new parameters, and calculate the probability, as in step 1



0.40

0.40

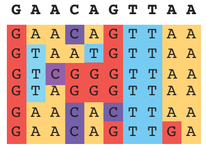
0.15

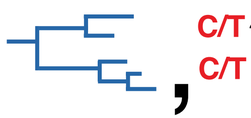
0.15

0.65

0.05

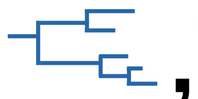
2500

Pr(

| 

, 

, Population size)

* priorPr(

, 

, Population size)

= 15

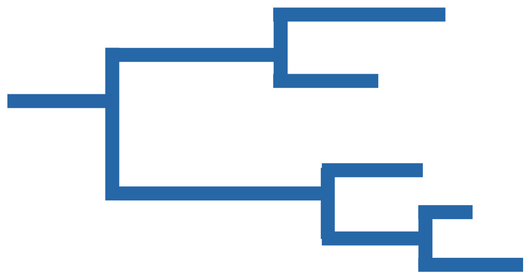
MCMC step 2b: compare that probability to the probability in step 1

If step 2's probability is higher, we accept the proposed parameters, and store them. If not, we reject with some probability.

Probability for step 1 = 10

Probability for step 2 = 15 -> **accept**

MCMC step 2b: compare the probability for step 2 to the probability for step 1



0.40

0.40

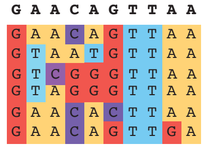
0.15

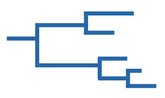
0.15

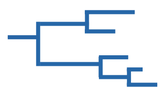
0.65

0.05

2500

Pr(

|  , C/T ⇝ A/G, C/T ⇜ A/G , Freq(A) Freq(C) Freq(G) Freq(T) , Population size)

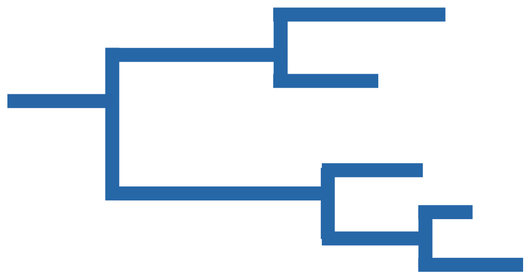
* priorPr( , C/T ⇝ A/G, C/T ⇜ A/G , Freq(A) Freq(C) Freq(G) Freq(T) , Population size)

= 15

accept

MCMC step 3: repeat for millions of steps

MCMC step 3: repeat for millions of steps



0.45

0.38

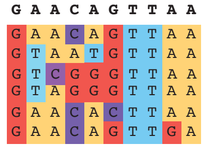
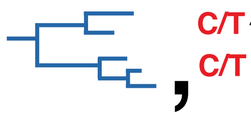
0.25

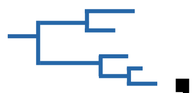
0.20

0.55

0.15

3000

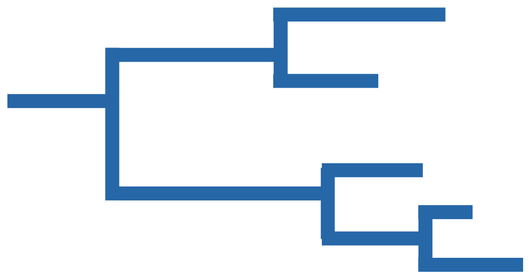
Pr( |  , C/T ⇝ A/G, C/T ⇜ A/G , Freq(A), Freq(C), Freq(G), Freq(T) , Population size)

* priorPr( , C/T ⇝ A/G, C/T ⇜ A/G , Freq(A), Freq(C), Freq(G), Freq(T) , Population size)

= 20

accept

MCMC step 4….



0.40

0.40

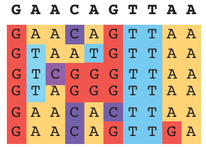
0.15

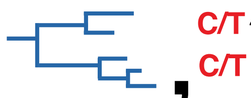
0.15

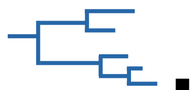
0.65

0.05

2500

Pr(

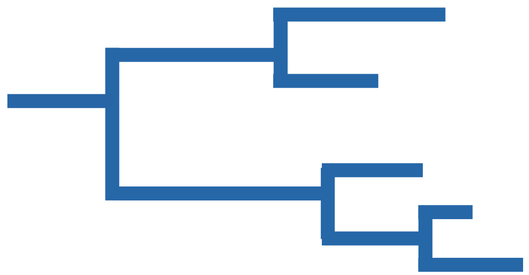
|  , C/T ⇝ A/G, C/T ⇜ A/G , Freq(A) Freq(C) Freq(G) Freq(T) , Population size)

* priorPr( , C/T ⇝ A/G, C/T ⇜ A/G , Freq(A) Freq(C) Freq(G) Freq(T) , Population size)

= 5

reject

MCMC step 5…



0.33

0.59

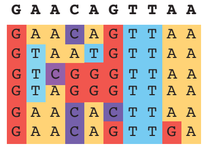
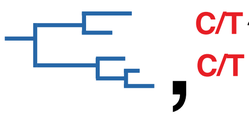
0.35

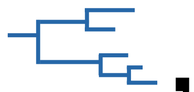
0.25

0.30

0.05

2500

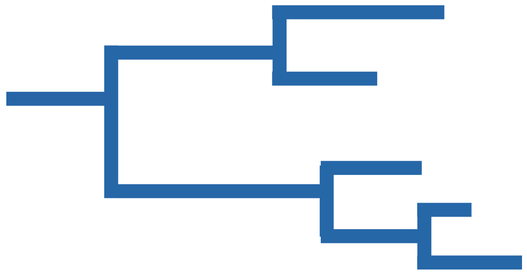
Pr( |  , C/T ⇄ A/G , C/T ⇄ A/G , Freq(A) Freq(C) Freq(G) Freq(T) , Population size)

* priorPr( , C/T ⇄ A/G , C/T ⇄ A/G , Freq(A) Freq(C) Freq(G) Freq(T) , Population size)

= 35

accept

MCMC step final: you've sampled a whole distribution of parameters, which are on average, more probable



C/T ⇄ A/G

C/T ⇄ A/G

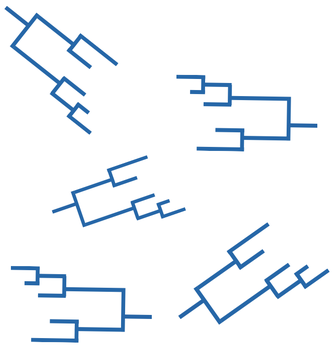
Freq(A)

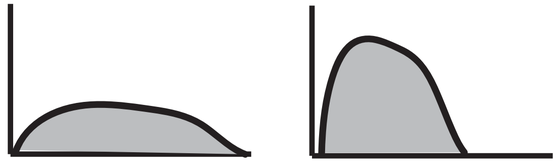
Freq(G)

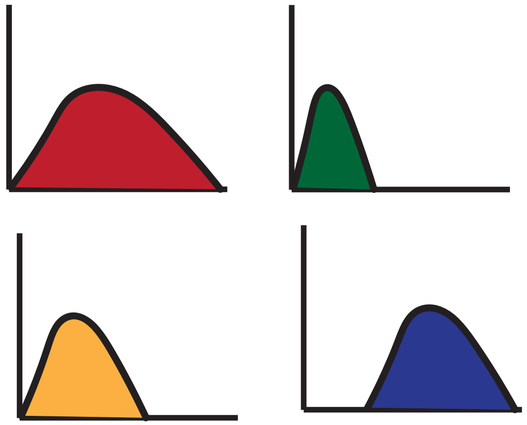
Freq(C)

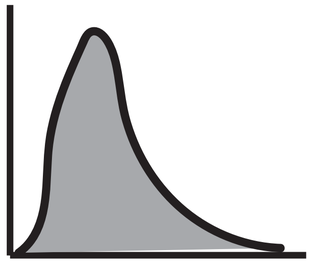
Freq(T)

Population size









Features of MCMC

1. Because the probabilities are multiplied by the prior, prior choice matters, and can influence your results.

2. MCMC chains are random. Usually, it takes the chain awhile to find a good search space. This initial phase is called “burnin” and is usually removed.

3. Because these chains are random, it is best to run multiple chains and compare results.

4. A good analysis is one that adequately explores parameter space, and converges. We evaluate “mixing” and “convergence” using Tracer.

Picking models and priors

Models:

1. What question are you trying to answer, and which model will help you answer it?
2. What biases exist in your data, and which model(s) will be most robust to those biases?

Picking models and priors

Models:

1. What question are you trying to answer, and which model will help you answer it?
2. What biases exist in your data, and which model(s) will be most robust to those biases?

Priors:

1. Generally want broad priors that allow sufficient space for exploration, but constrain the analysis to a reasonable array of values.
2. Best to think about which distributions reasonably match the estimate you are trying to make.
3. If model and data are particularly sensitive to a prior, good to evaluate results across multiple values of that prior.

BEAST vs. Nextstrain

	Nextstrain	BEAST Bayesian Evolutionary Analysis Sampling Trees
Tree inference method	Maximum likelihood, or approximate likelihood; results in single, best tree	Bayesian inference with MCMC, results in a posterior set of trees
Interactive visualization?	Yes	No
Substitution models?	Yes	Yes
Time-resolved trees?	Yes, inferred with TreeTime under coalescent model	Yes, inferred by MCMC, using a large variety of models
Flexible demographic models?	Yes, skyline or constant	Yes, an enormous array of models
Non-coalescent models?	No	Yes
Runtime	Fast	Slow
Input requirements/decisions	Un-aligned sequences; must decide subsampling regime, tree inference method, substitution model	Aligned, subsampled sequences; must decide on tree model, clock model, priors, chain length

Extra resources:

- Felsenstein's pruning algorithm: https://link.springer.com/article/10.1007/BF01734359

- MCMC robot example and exploration from Lewis lab: https://plewis.github.io/applets/mcmc-robot/

BEAST Lab

BEAST

Bayesian Evolutionary Analysis Sampling Trees

1. How did H3N8 transmit between hosts and geographic locations?
2. How did the H3N8 population size change during transmission in horses and dogs?
3. Is there evidence that H3N8 evolved at different rates in horses and dogs?

Some common prior distributions:

Markov chain Monte Carlo

Step 1:

1. Randomly pick starting values for all parameters you would like to estimate.
2. Evaluate the likelihood of those parameters given the data.
3. Multiply that likelihood by the prior (since we're doing this Bayesian!)
4. Store that probability.

Step 2:

5. Repeat steps 1-4.
6. Compare the probability from step 1 to step 2. If the parameters result in a higher probability for step 2 than step 1, accept step 2, store the values, and continue your search from there. If the probability is lower, reject the step.

Steps 3-end

7. Repeat this process, saving more probable steps and rejecting less probable ones, until the chain converges (10,000-100,000 steps)