

Phylogenetics and Coalescent Processes

Phylogenetics

- If our observations are different species, then we are primarily interested in speciation events
- Can ask several questions
 - what is the “true” tree
 - when did the speciation events occur?
 - are speciation events correlated?

Coalescence

- We might be interested in individuals within a species
- Now we are not so interested in when individual lines separated (“coalesced”), but rather the overall pattern
- Interested in the dynamics of the process
 - time to Most Recent Common Ancestor
 - rates of migration

Phylogenetics: Species Trees

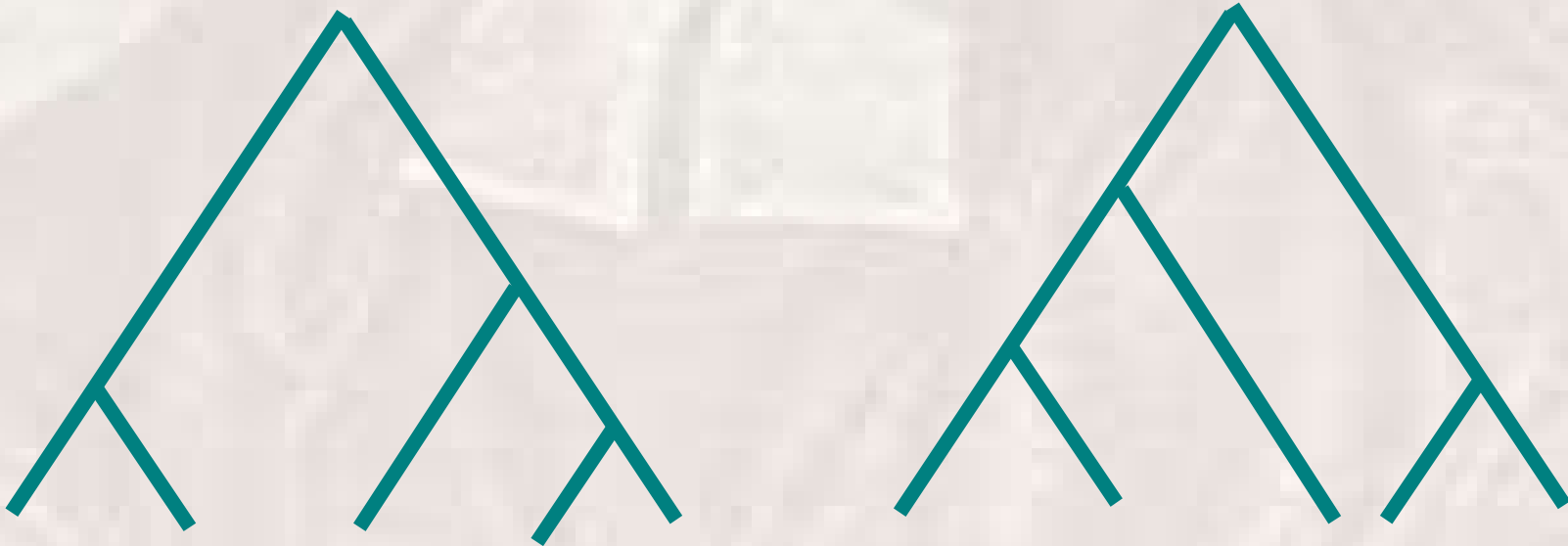
- Get sequences from several species
- Want to reconstruct the species tree
- Need two models:
 - tree generation process
 - speciation and extinction
 - sequence evolution
 - conditioned on the tree
- Several possible models for both

Three Aspects of Trees

- Topology
 - tree shape
- Times of events speciation
 - branch lengths
- Both depend on the process of species evolution
- Sequence evolution on the tree
 - assume a neutral model

Topology

- The shape of a tree
- These have the same topology:



- We can stretch and rotate them until they are identical

Models

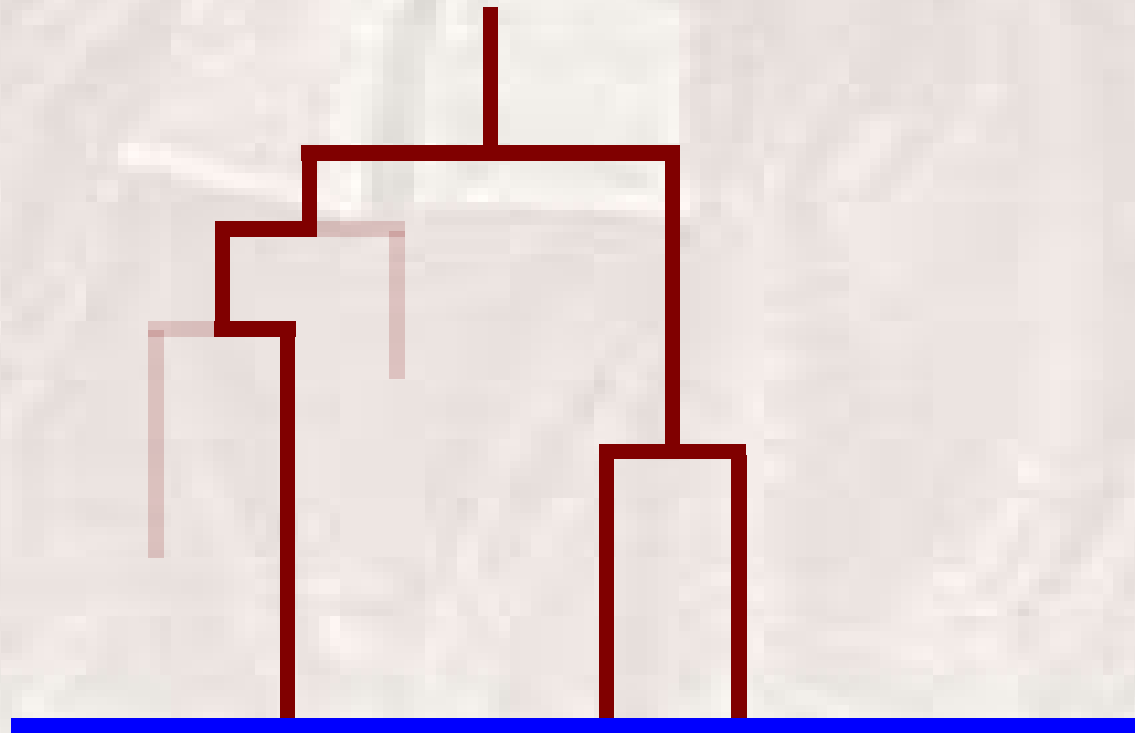
- We need a model of the process of creating a tree
- Simple model: birth-death process
 - classic model from stochastic processes
- Birth: speciation
- Death: extinction
- Can make complex
 - here we will only look at a simple one

Birth-Death Processes

- Events occur at a constant rate and are independent
- Start with a single lineage
- In a time interval Δt , for one lineage:
 - Birth rate $\lambda\Delta t$
 - Death rate $\mu\Delta t$
- A birth: a lineage splits into two
- Death: the lineage stops

What we See

- For a present day sample, we do not see the species that have gone extinct



Modelling the Tree

- We are interested in the tree for the species we observe
- The tree can be split into two parts:
 - τ : the topology
 - the shapes of births and deaths
 - t : the speciation times
 - order of events
- Every topology is equally likely
- Branch lengths depend on the birth and death rates

Nice Equations

- Assuming we see all s species the joint density is:

$$f(t, \tau | \mu, \lambda) = \frac{2^{s-1} \mu^{s-2} \prod_{i=2}^{s-1} p_i(t_i)}{[p_0(1)]^{s-2} s! (s-1)}$$

- where $p_i(t)$ is the probability that a lineage leaves i descendants after time t , and is:

$$p_i(t) = \left(\frac{\lambda}{\mu} \right)^i \frac{(\lambda - \mu)^2 e^{-(\lambda - \mu)t}}{(\lambda - \mu e^{-(\lambda - \mu)t})^2} \left(\frac{\mu(1 - e^{-(\lambda - \mu)t})}{\lambda - e^{-(\lambda - \mu)t}} \right)^{i-1}$$

How We Get the Structure

- Because each tree is equally likely, we need some other information to compare trees
- This is where we use the sequence data
- If we have a tree (both τ and t) we can calculate the likelihood of any sequence
- We need a mutation model

Mutation Models

- Each site has one of 4 bases (A, T, G, C)
- Write the rate of mutation as a matrix, **Q**:

		To			
		A	T	G	C
From	A	-pAA	pAT	pAG	pAC
	T	pTA	-pTT	pTG	pTC
	G	pGA	pGT	-pGG	pGC
	C	pCA	pCT	pCG	-pCC

- Technical point: rows sum to 0
 - so $pAA = pAT + pAG + pAC$ etc.

Simple Mutation Model

- Every time there is a mutation, the base change is to each one of the others with equal probability

– Jukes-Cantor

		To			
		A	T	G	C
From	A	-3	1	1	1
	T	1	-3	1	1
	G	1	1	-3	1
	C	1	1	1	-3

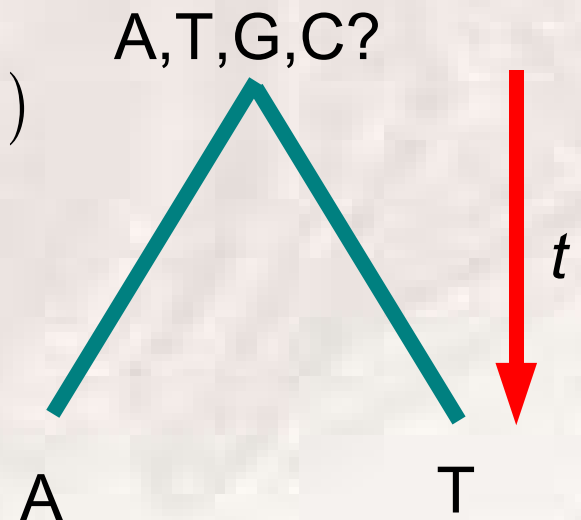
- Lot of alternative models
 - see a phylogenetics course!

What we do with it

- Q gives the rate of substitution
- If π_t is the probability density for the state of a site at time t , then $\pi_{t+s} = \pi_t e^{Qt}$
- If we observe two species, and one site:

$$P(X|t) = \sum_{i=A,T,G,C} P(i \text{ to } A|t) \cdot P(i \text{ to } T|t)$$

- Read $P(i \text{ to } A | t)$ from $\pi_t e^{Qt}$



Full Sequence Likelihood

- The likelihood for the full sequences (\mathbf{X}) is:

$$f(\mathbf{X} | \tau, \mathbf{t}, m, \kappa) = \prod_{j=1}^n f(\mathbf{X}_j | \tau, \mathbf{t}, \mathbf{Q})$$

- $f(\mathbf{X}_j | \tau, \mathbf{t}, \mathbf{Q})$ is the likelihood for the bases seen in the sample \mathbf{t} one site
- Assume independence between sites
- Depends on the tree structure (τ, \mathbf{t}) and the mutation model (\mathbf{Q})

Putting Them Together

- We have the following likelihoods

- The tree: $f(t, \tau | \mu, \lambda)$

- The sequences: $f(\mathbf{X} | t, \tau, \mathbf{Q})$

- So the full likelihood is

$$f(\mathbf{X} | \mu, \lambda, \mathbf{Q}) = f(\mathbf{X} | t, \tau, \mathbf{Q}) f(t, \tau | \mu, \lambda)$$

- To get the posterior, we just multiply this by priors for μ , λ and \mathbf{Q}
- And then read off the posterior for τ
 - the probabilities of each tree

The Practicalities

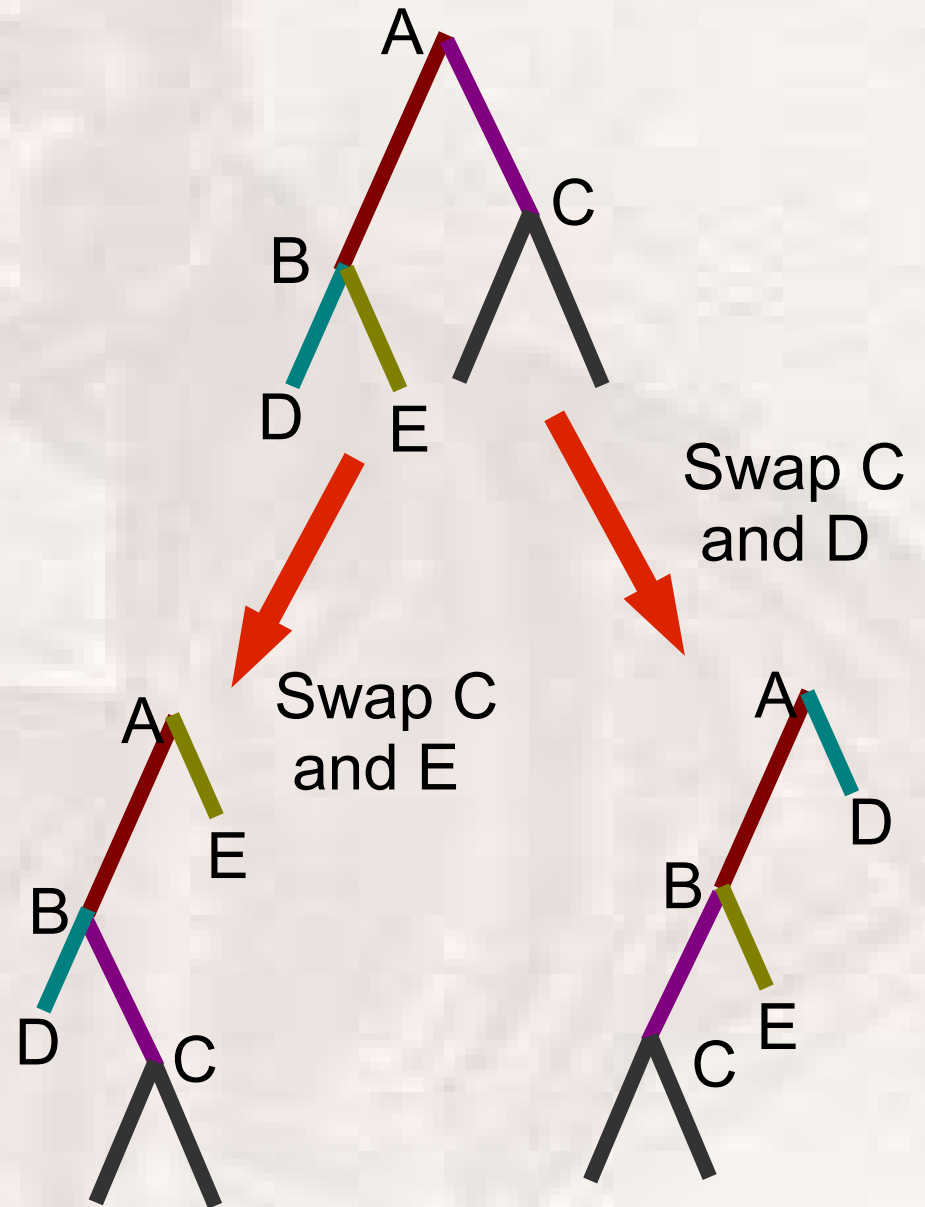
- Update by MCMC
- Allowed to update parameters one at a time
 - use the Metropolis-Hastings algorithm
- Can update in two stages
 - parameters, given the present tree
 - the tree, given the parameters
- Updating parameters is like updating any other parameters
- Tree more difficult

Updating The Tree

- If there are few species, can go through every tree and estimate the other parameters
 - compare their marginal posteriors
- With most problems there are too many possible trees
- Instead, use Metropolis-Hastings to sample the trees from the posterior
- Have to find a way of proposing a new tree

One Proposal

- Nearest Neighbour Interchange
- Take a branch
- Swap two of its three daughters
- Can do in two ways
- Look at all possible swaps
- Each one has same probability



Putting it all together

- Yang & Rannala (1997)
- More complicated mutation model
- Update tree topologies with M-H
- For each topology, use importance weighting to estimate parameters' distributions
 - including the branch lengths
 - get the posterior probability for the tree
 - Do for every tree proposed, so for each tree you only need to do it once

Some Data

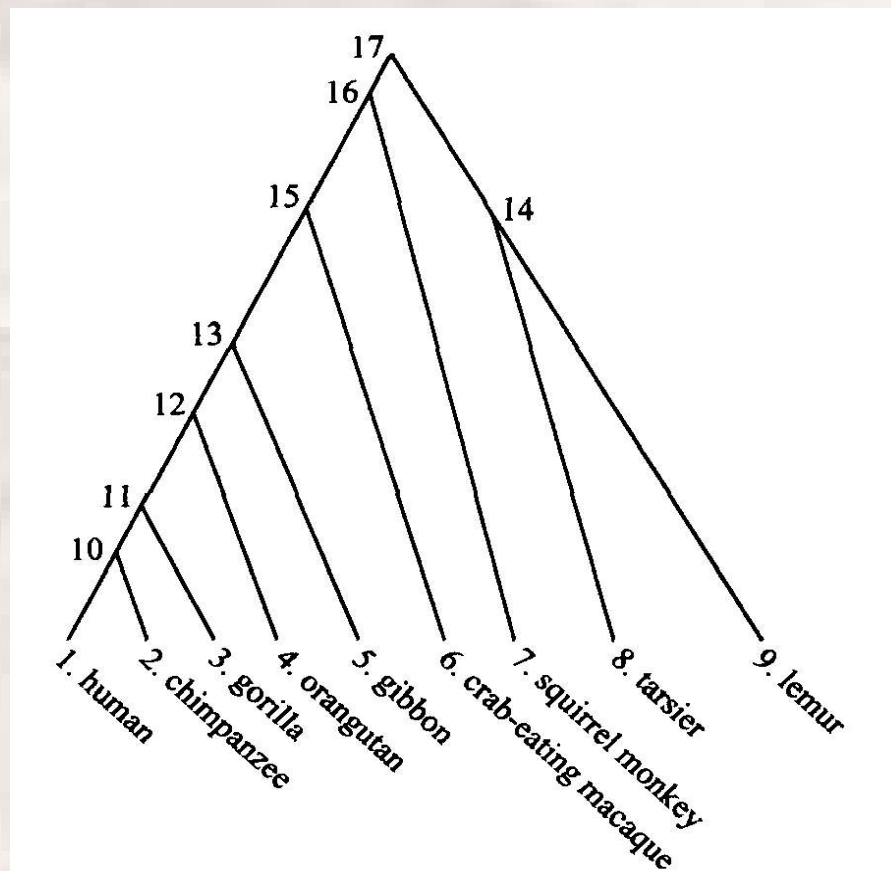
- As an example, use primates
 - 10 species
 - 2 protein coding sequences, 3 tRNAs
- Look at posteriors for trees
 - marginalise over parameters
- See both the topology and the branch lengths
 - timing of speciation

What They Get

- Data supports one topology
- Only differences in time of node 14
 - between 13 and 17

Labeled History	Tree Topology	HBA ($S = 185$)	
		ℓ	π
1	(((((((12)3)4)5)6)7)(98))	-5,261.89	0.745
2	(((((((12)3)4)5)6)7)(98))	-5,263.17	0.208
3	(((((((12)3)4)5)6)7)(98))	-5,268.03	0.002
4	(((((((12)3)4)5)6)((98)7))	-5,265.07	0.031
5	(((((((12)3)4)5)6)((98)7))	-5,269.86	0.000
6	(((((((12)3)4)5)6)((98)7))	-5,269.21	0.001
7	(((((((1(23))4)5)6)7)(89))	-5,268.05	0.002
8	(((((((1(23))4)5)6)7)(89))	-5,269.67	0.000
9	(((((((12)3)4)5)6)(98))7)	-5,266.15	0.011
10	(((((((12)3)4)5)6)(98))7)	-5,270.25	0.000

Posterior Probabilities



Coalescence Processes

- Trees within a species
- e.g. mitochondrial DNA
- Consider we have a sample from a population
- Interested in what the genealogy of the genes can tell us
- Depends on the population dynamics
 - e.g. population size

The Population Model

- We have a population of constant size, and follow a single gene
- Variation in the number of offspring means that some lineages die out
 - average offspring = 1, so any variation means there must be some zeroes
- The branching process is more complicated because branches are not independent
 - if one wins, another loses

The Coalescent

- When we look at a sample now, we cannot see the lineages that are extinct
- Kingman pointed out that we do not need to see them to model the population
- If we follow the lineages of our sample back, then we will get a tree
- Crucially, we can model that process in terms of the population dynamics

Look Back in...

- The innovation:
- Reverse time, and talk about lineages joining
 - coalescing
- This reversed process has nice mathematical properties
 - times to coalescence are just exponential distributions
 - process has no memory
 - times measured as population size
 - rate of coalescence depends on population size

The Tree's Prior

- All the coalescent times are independent
- All pairs of branches equally likely to coalesce
- Joint probability of times is:

$$P(t_1, t_2, \dots, t_{n-1}) = \prod_{i=1}^{n-1} \binom{n+1-i}{2} \exp\left(-\binom{n+1-i}{2} t_i\right)$$

- All pairs of branches equally likely to coalesce, so this is proportional to the probability density of any tree

The Estimation

- We can use molecular data to make inferences about the coalescent
 - just as before!
- The tree itself is not so important
 - who cares if individuals 1402 and 1839 are distantly related?
- Instead we look at the parameters of the process
 - e.g. time to **M**ost **R**ecent **C**ommon **A**ncestor

Improvements in Method

- We could estimate the probabilities of the different states using pruning
- Instead, we can use data augmentation
 - estimate the states of all the nodes
- Then calculating the likelihood is easier
 - just multiply the likelihoods for each branch
- Cost: augmenting the data takes time
 - but may still be quicker

The Data

- Our example: Wilson & Balding, 1998
- We use genetic data again
- Can use sequence data, model will be the same as above
- Here we will use microsatellite data
 - interesting mutation model
 - Five microsatellites, human Y chromosome
 - Mutation model: Stepwise Mutation Model

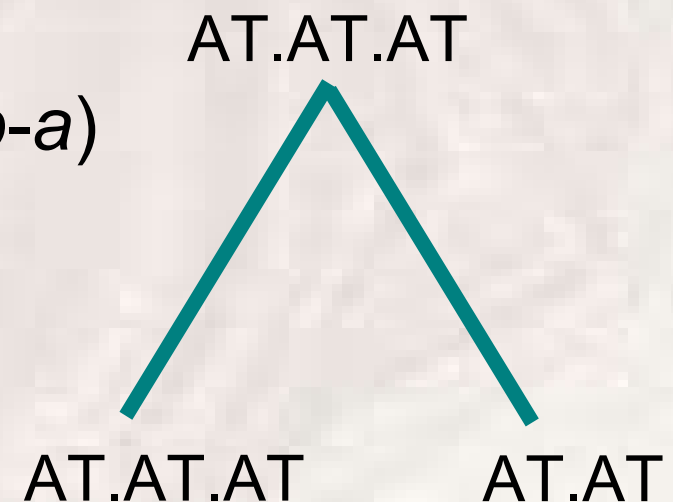
SMM: the Equation

- We want to know the likelihood of going from the state at the start of a branch to the state at the end
- If we want to go from a to b repeats, we calculate all possible ways

– all ways in which up – down = $(b-a)$

- This is:

$$v_d(t, \theta) = e^{t\theta/2} \sum_{k=0}^{\infty} \frac{(t\theta/4)^{2k+d}}{k!(k+d)!}$$



Updating the Tree

- The tree can be updated in the same way as before
- Wilson & Balding use a different approach
- Chose an internal node at random and try to move it to another part of the tree
- Choose to move it to a place with a similar genotype with greater probability
 - allows large changes in topology
 - works OK in practice

Stages of the algorithm

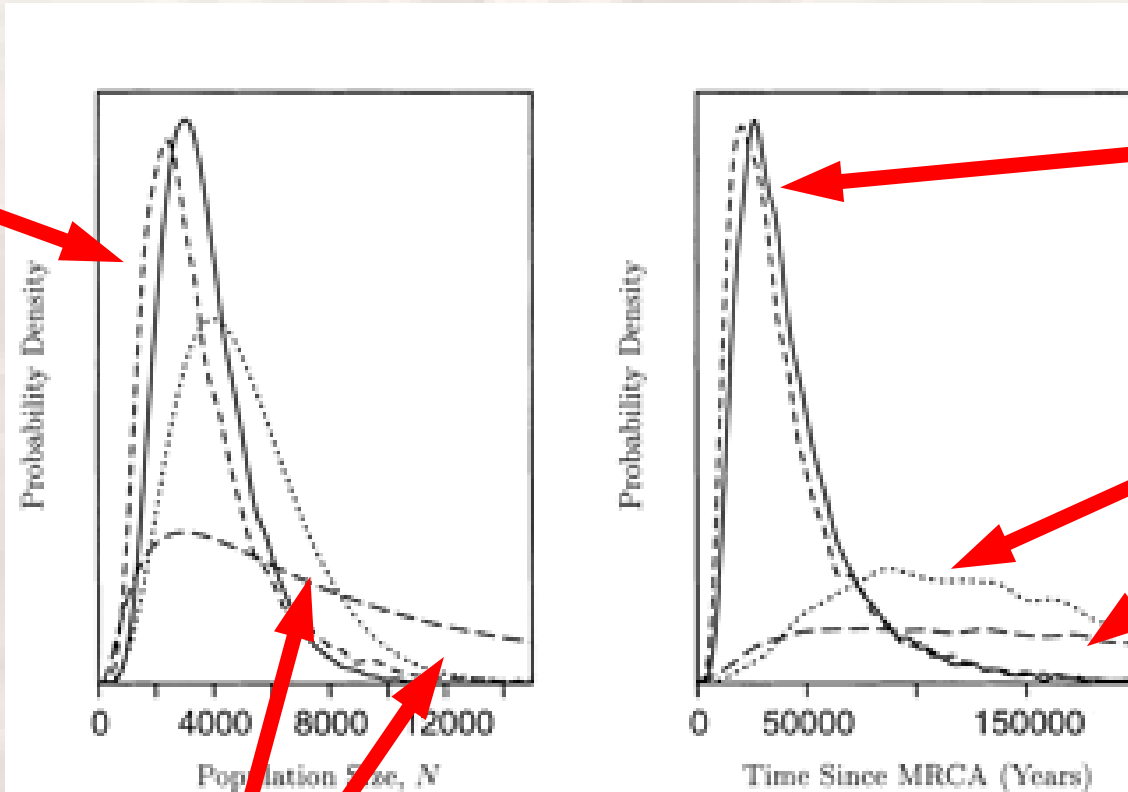
- Update the tree topology
 - swap branches
 - if swap, determine new branch lengths
 - i.e. where the new branch is
- Augment internal nodes
 - M-H again
- Update branch lengths
- Update parameters
 - population size, mutation rate

The Results

- Interested in time to MRCA
- Is it close to 200 000 years?
 - the “out of Africa” hypothesis
- Also get posterior for effective population size

Graphs

Posteriors



Posteriors

Priors

Priors

Coalescence and Phylogeny

- The tools for analysing coalescent and phylogenetic processes are similar
 - fit a model to the tree
 - estimate the tree topology and the parameters of the model
- But, here they have been used for different things
 - phylogenetics: what is the “true” tree
 - coalescence: what caused the tree?

Summary...

- The estimation proceeds along similar lines in both cases
 - estimate the tree topology, branch lengths, mutation process
- Differences due to the data or technique
 - type of marker
 - updating methods evolve
 - e.g. data augmentation used because it is faster
- Both methods give a tree and the parameters

Why Bayesian is Good

- Both methods give a tree and the parameters
- Because we are Bayesian, we can then take the marginal distribution for what we are interested in
- Phylogenetics: marginal for the tree
 - include uncertainty in the parameters
- Coalescence: marginal for parameters
 - include uncertainty in the tree
- Could also use to predict ancestral states
 - marginalise over tree and parameters