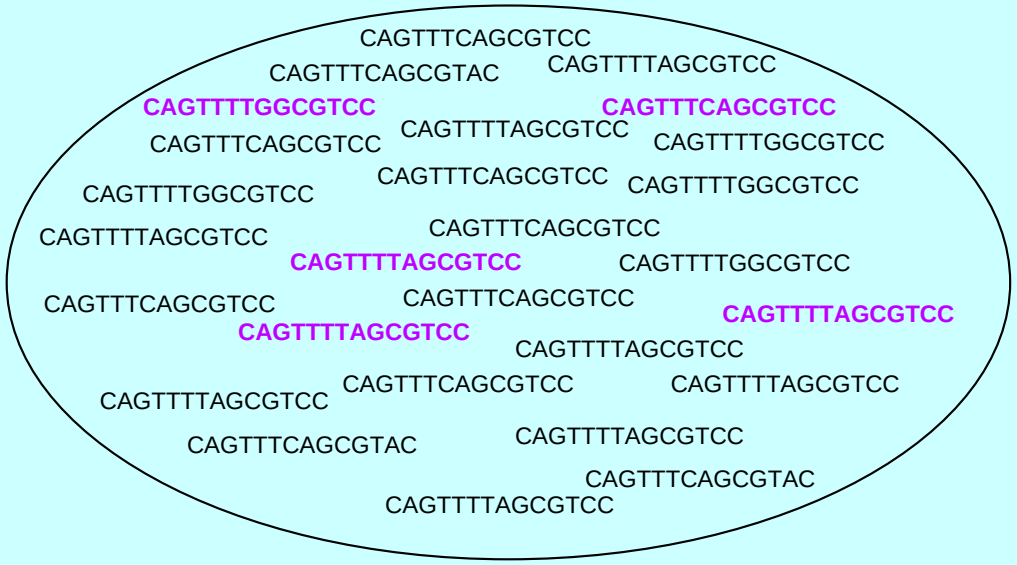
Lecture 8. Coalescent likelihoods and introduction to MCMC

Joe Felsenstein

Department of Genome Sciences and Department of Biology

Some typical data with within-population variation

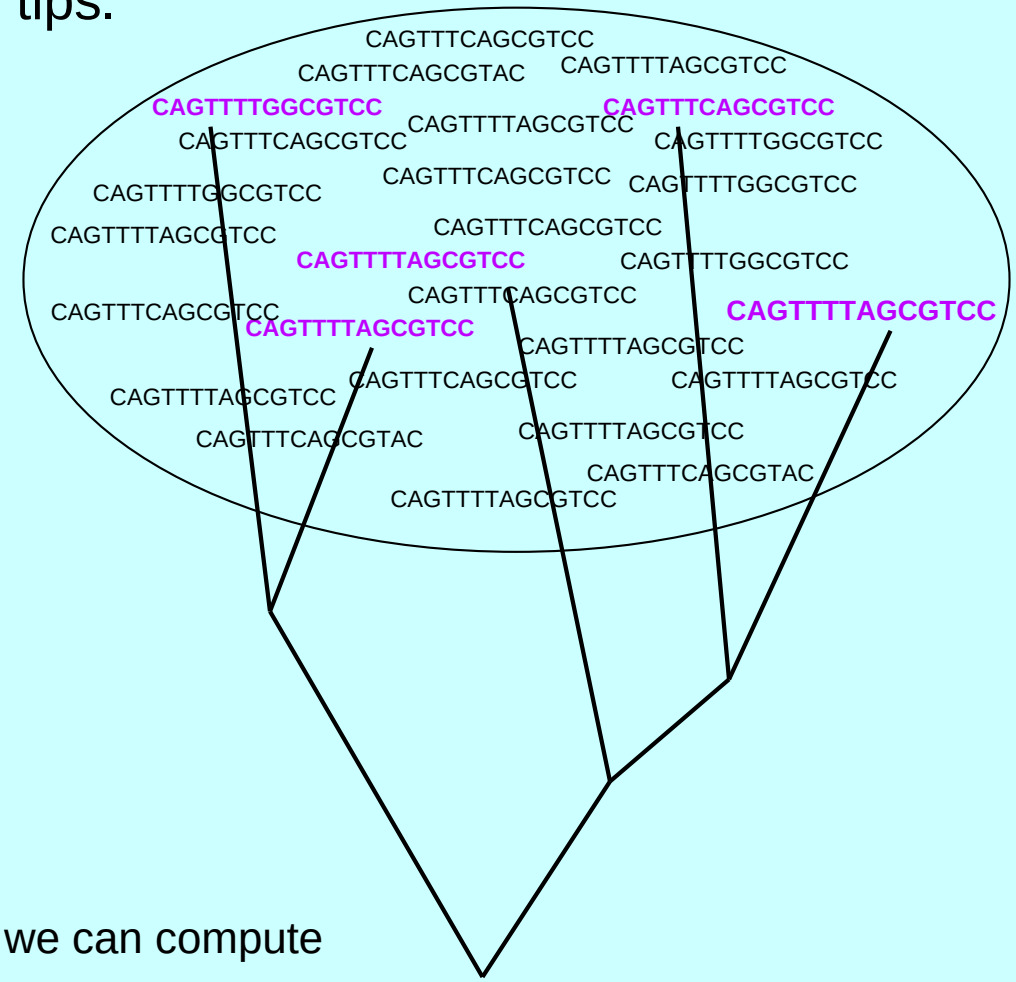


L = Prob (CAGTTTCAGCGTCC , CAGTTTCAGCGTCC , ...) = ??

To infer parameters of evolutionary-genetic models, we need to compute the likelihood for a set of genotypes sampled from a population. With few exceptions, no expressions for this likelihood exist.

If we knew the genealogical tree, we could ...

... as we know from work on phylogenies how to compute the joint probability at the tips.



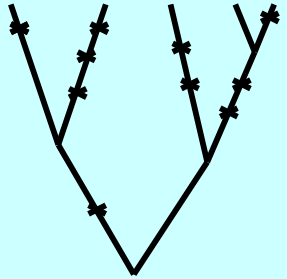
so we can compute

Prob(CAGTTTCAGCGTCC , CAGTTTCAGCGTCC , ... | Genealogy)

but how to compute the overall likelihood from this?

Two sources of variation

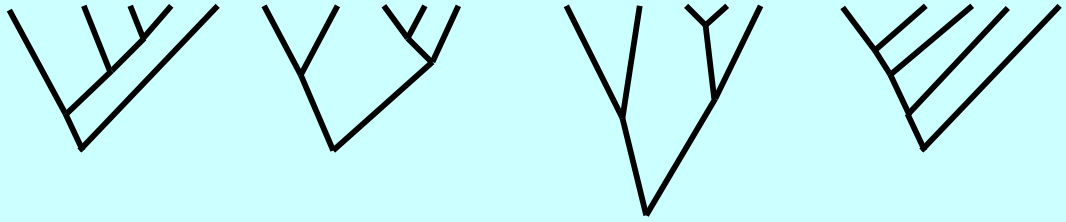
(1) Randomness of mutation



affected by the mutation rate u

can reduce variance of number of mutations per site per branch by examining more sites

(2) Randomness of coalescence of lineages



affected by effective population size N_e

coalescence times allow estimation of N_e

can reduce variability by looking at

(i) more gene copies, or

(ii) more loci

The basic equation for coalescent likelihoods

In the case of a single population with parameters

$\mathsf{N_e}$ effective population size

μ mutation rate per site

and assuming G' stands for a coalescent genealogy and D for the sequences,

$$\mathsf{L} = \text{Prob}\,(\mathsf{D} \,|\, \mathsf{N_e}, \mu)$$

$$= \sum_{\mathsf{G}'} \underbrace{\text{Prob}\,(\mathsf{G}' \,|\, \mathsf{N_e})}_{\text{Kingman's prior}} \underbrace{\text{Prob}\,(\mathsf{D} \,|\, \mathsf{G}', \mu)}_{\text{likelihood of tree}}$$

Rescaling branch lengths ...

Rescaling branch lengths of G' so that branches are given in expected mutations per site, $G = \mu G'$, we get (if we let $\Theta = 4N_e\mu$)

$$L = \sum_{G} \text{Prob}\,(G \mid \Theta)\,\text{Prob}\,(D \mid G)$$

as the fundamental equation. For more complex population scenarios one simply replaces Θ with a vector of parameters.

In a simple example we can compute likelihood curve

If two sequences of length 1000 differ by 0.5% under a Jukes-Cantor model

$$\mathsf{L}(\mathsf{N},\mu) \quad = \text{ Prob } (5 \text{ different} \mid \mu, \mathsf{t})$$

$$= \quad \int_0^\infty \tfrac{1}{2\mathsf{N}}\, \mathsf{e}^{-\frac{\mathsf{t}}{2\mathsf{N}}} \binom{1000}{5} \left(\tfrac{3}{4}\left(1 - \mathsf{e}^{-\frac{4}{3}\mu(2\mathsf{t})}\right)\right)^5 \left(\tfrac{1}{4}\left(1 + 3\,\mathsf{e}^{-\frac{4}{3}\mu(2\mathsf{t})}\right)\right)^{995} \mathsf{dt}$$

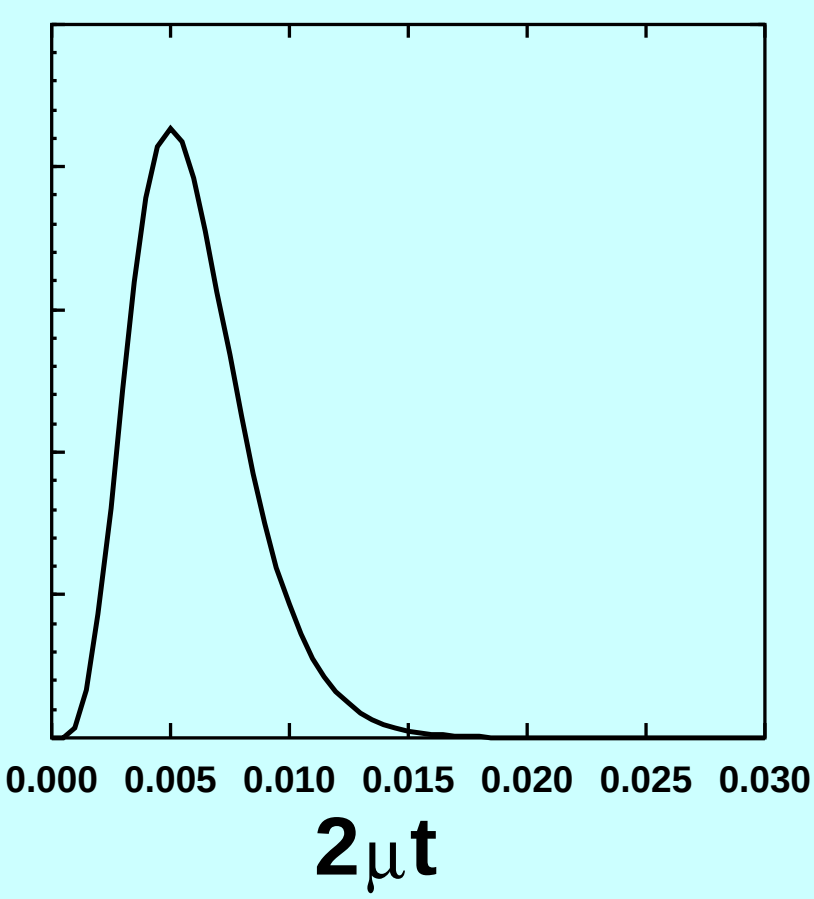
In the product, the left two terms are the exponential density from the Kingman coalescent time density, the right three terms are the binomial probability which is the likelihood for a tree whose branch length is $2\mu\mathsf{t}$.

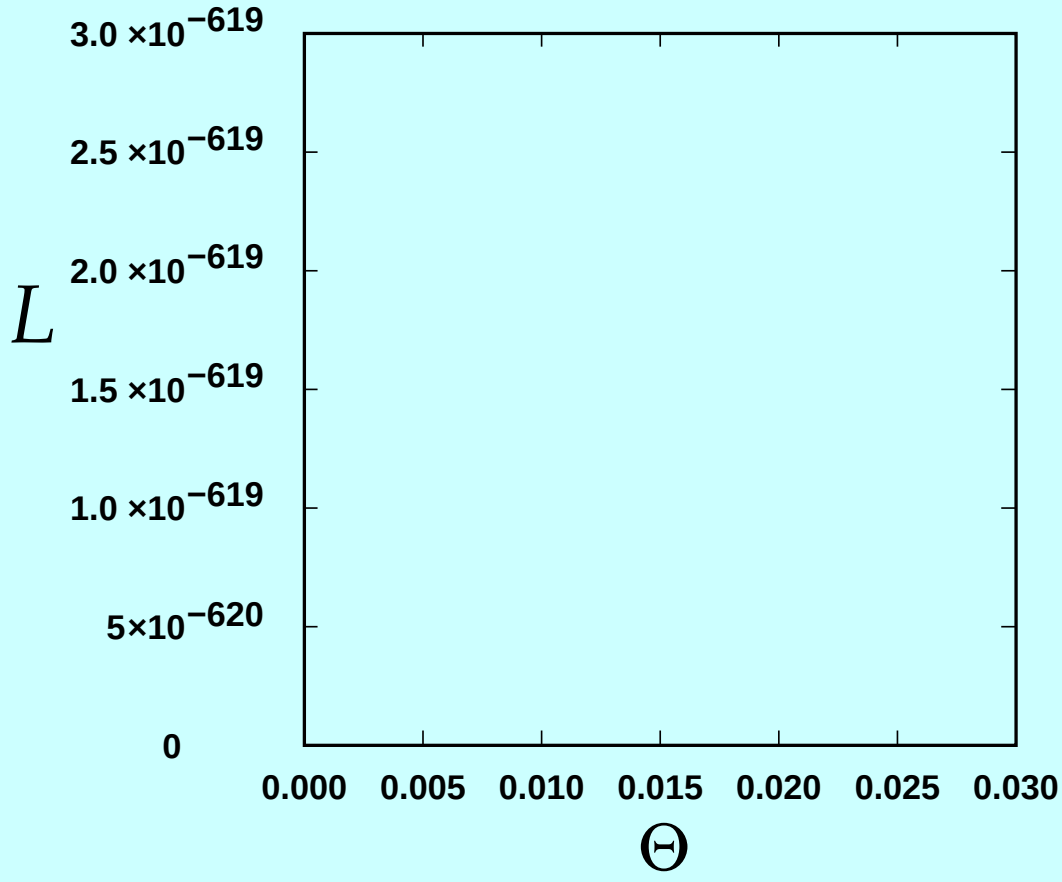
Replacing $\mu\mathsf{t}$ by a new variable u, this is a function only of $4N\mu$

$$\int_0^\infty \frac{2}{4\mathsf{N}\mu}\, \mathsf{e}^{-\frac{2\mathsf{u}}{4\mathsf{N}\mu}} \binom{1000}{5} \left(\frac{3}{4}\left(1 - \mathsf{e}^{-\frac{8}{3}\mathsf{u})}\right)\right)^5 \left(\frac{1}{4}\left(1 + 3\,\mathsf{e}^{-\frac{8}{3}\mathsf{u}}\right)\right)^{995} \mathsf{du}$$

and can be easily evaluated numerically to get $\mathsf{L}(4\mathsf{N}\mu)$.

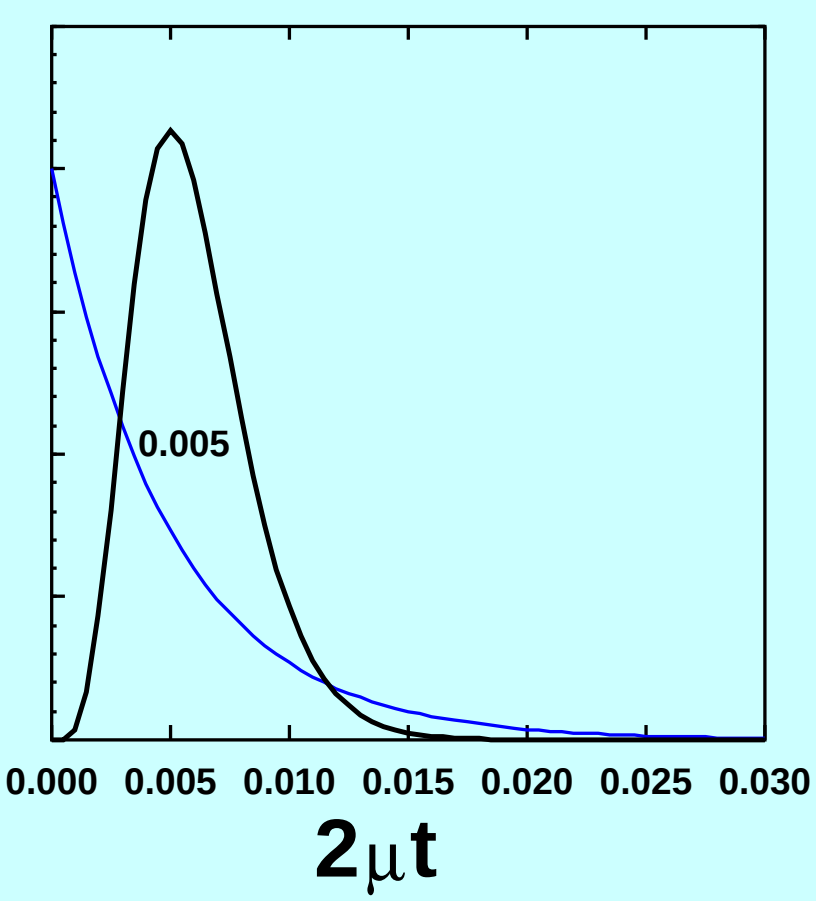
If two sequences with 1000 bases 0.5% different

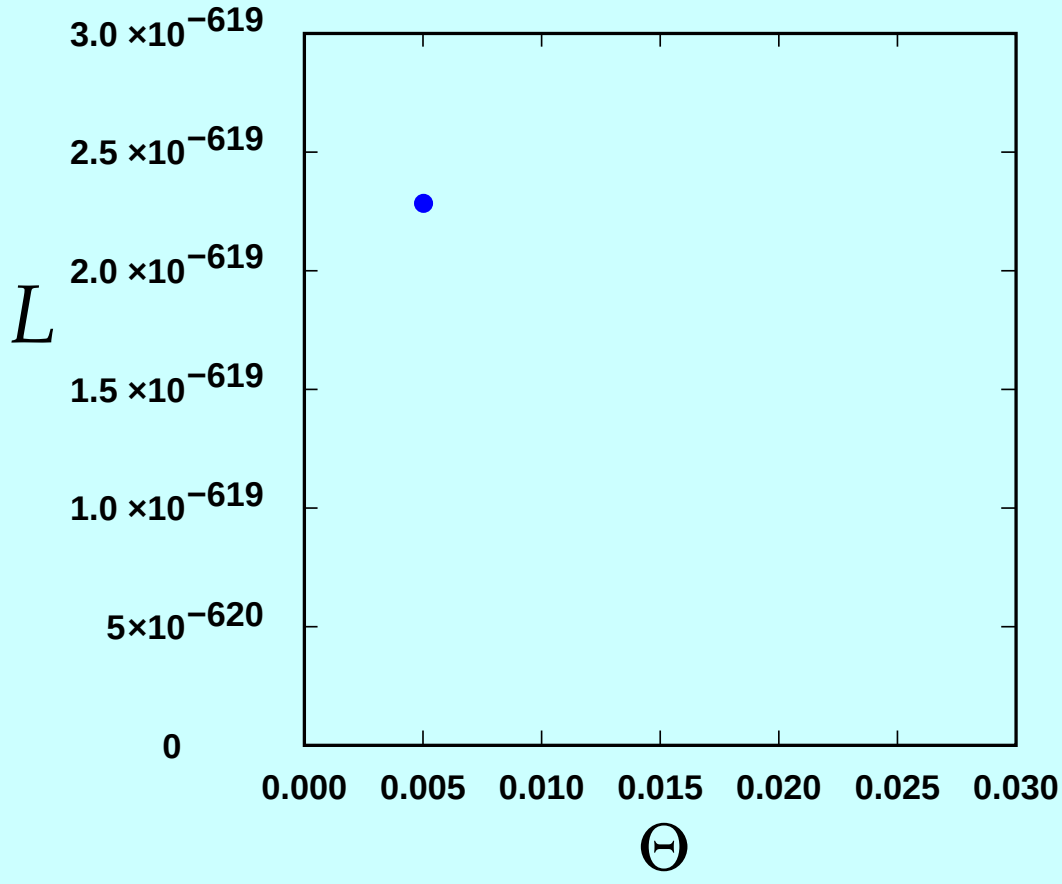




This is the likelihood (right) term in the preceding equation, the binomial probability that 5 sites would differ out of 1000, as a function of the branch length $2\mu t$

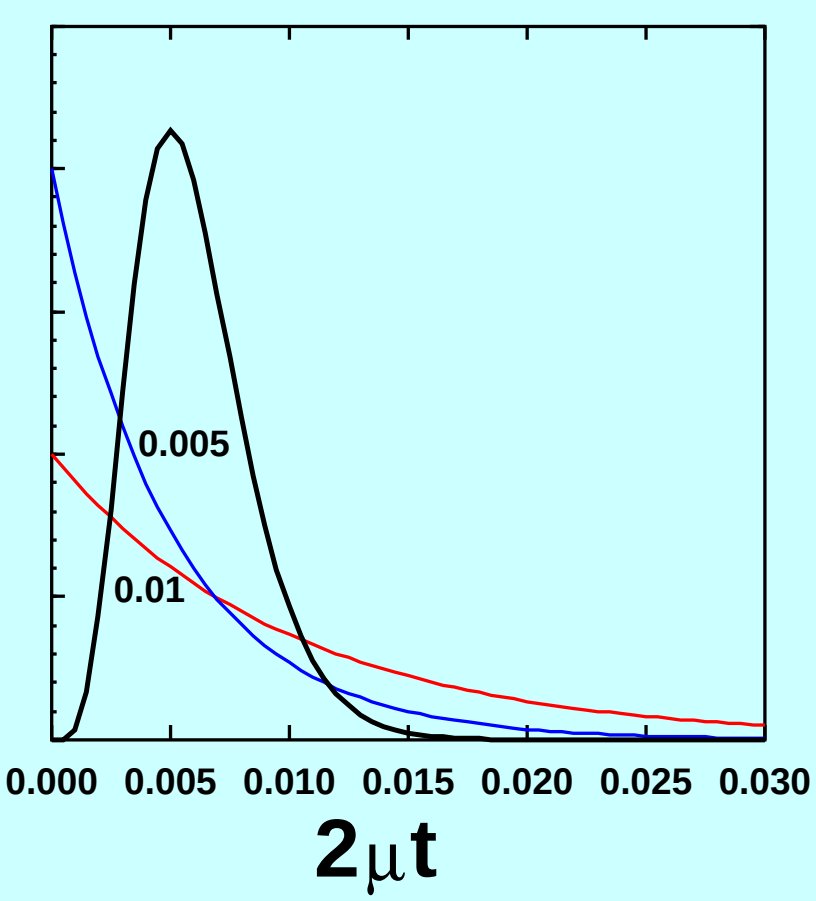
If two sequences with 1000 bases 0.5% different

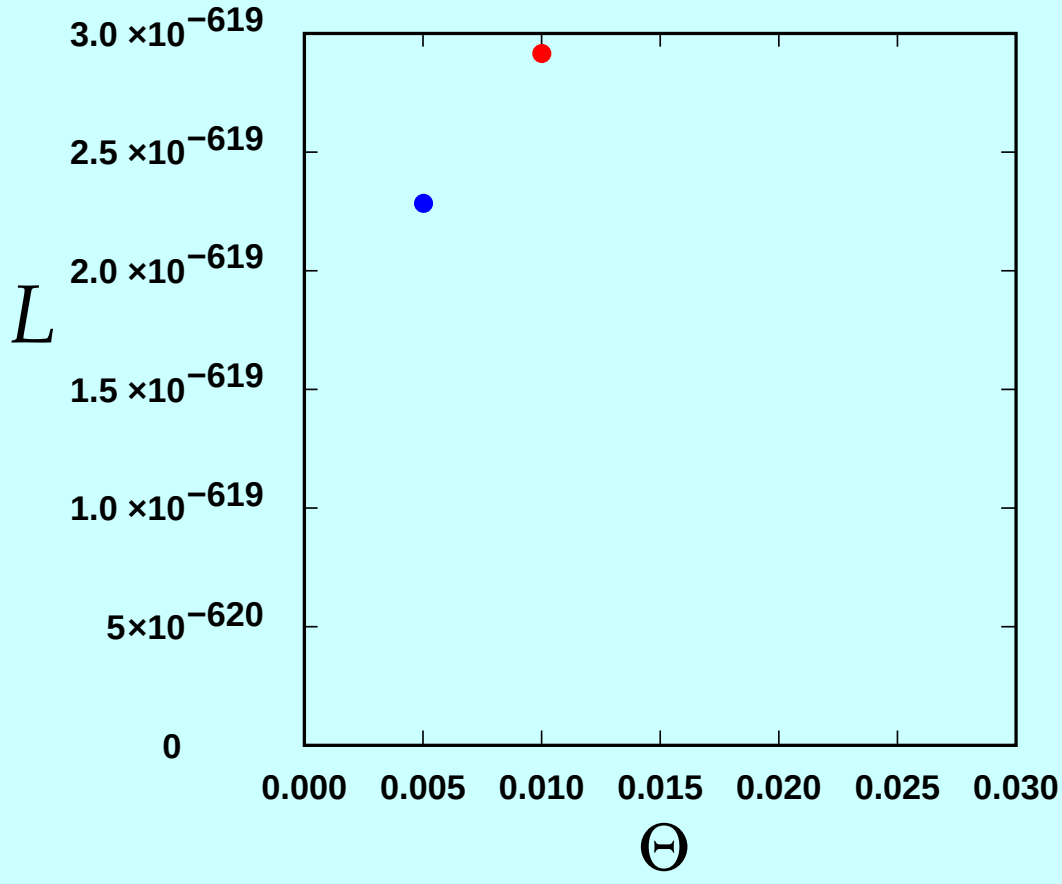




For a given value of Θ, 0.005, we plot the two-tip Kingman coalescent density function of the branch length (twice the time to coalescence), and we can integrate the product of that with the likelihood curve. The right graph shows that integral plotted against Θ.

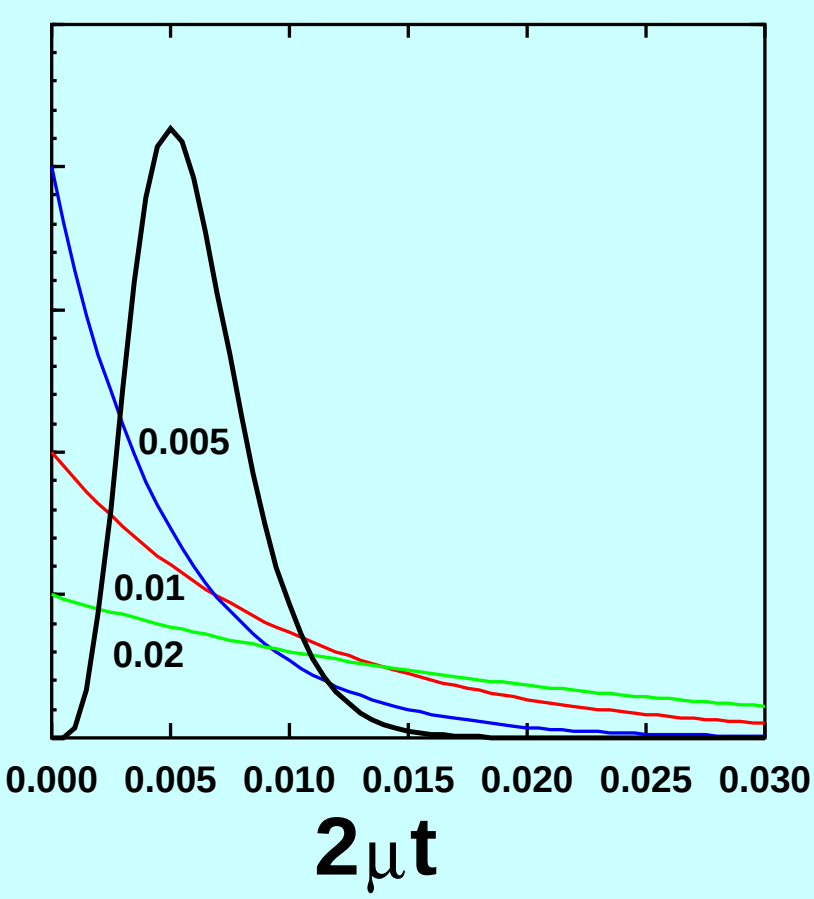
If two sequences with 1000 bases 0.5% different

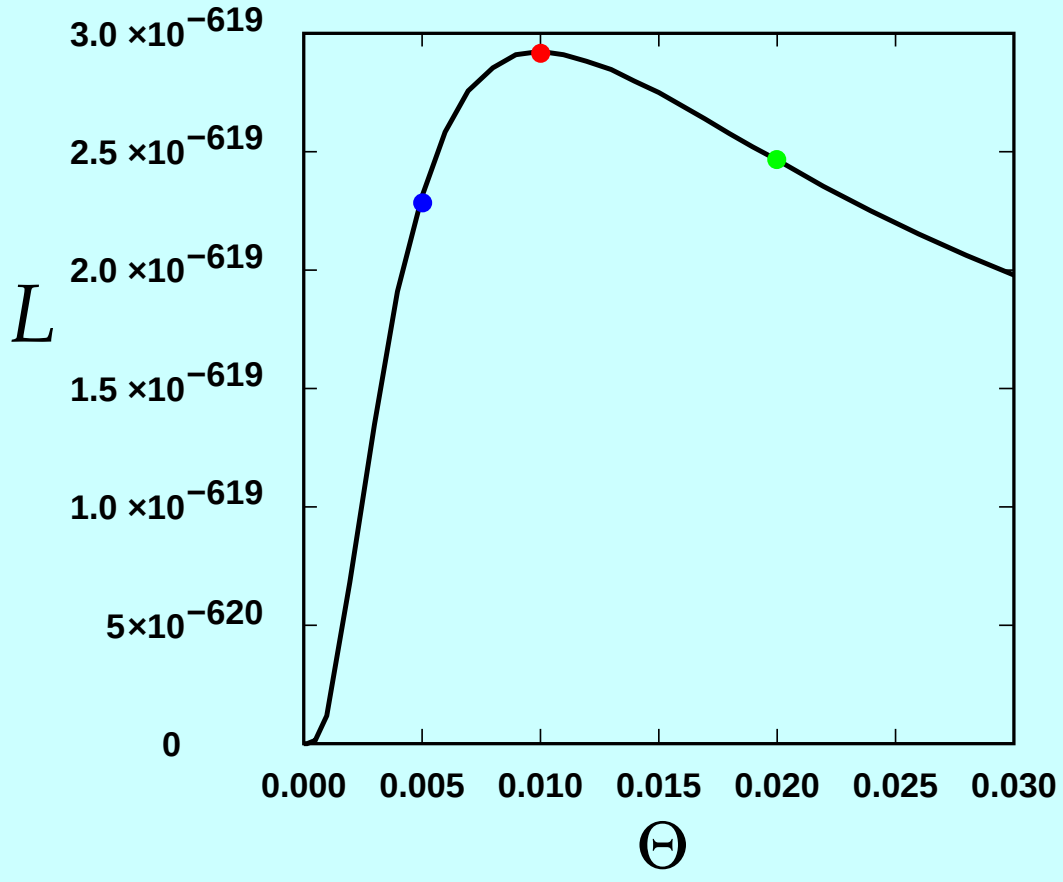




On the right is the same integral, but now using instead the Kingman coalescent density for $\Theta = 0.01$.

If two sequences with 1000 bases 0.5% different





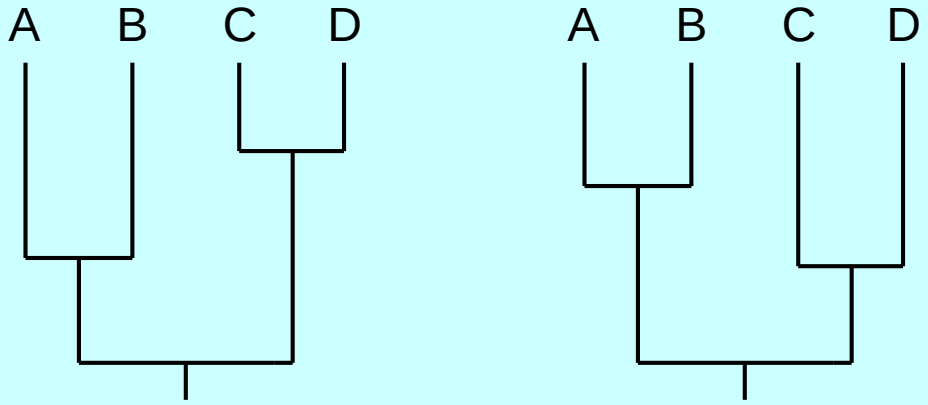
Finally, we can do it also for $\Theta = 0.02$, and connect the three points in the right-hand graph by a curve (this graph shows the curve we would get if we did a finer grid of points).

Labelled histories

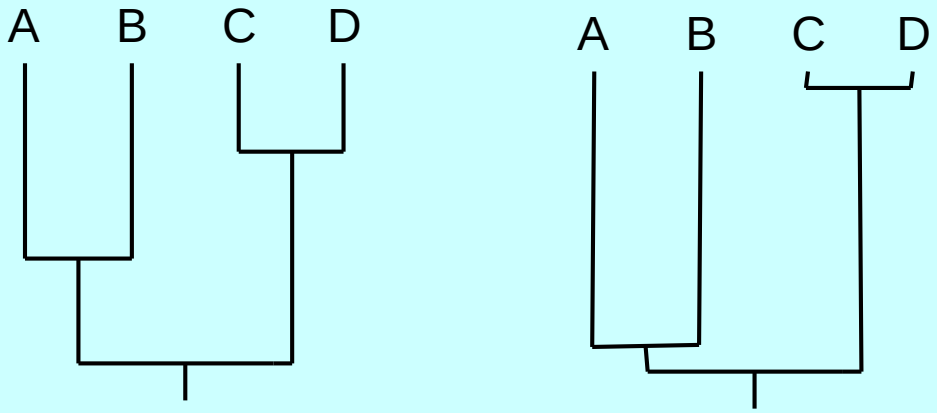
(Edwards, 1970; Harding, 1971)

Trees that differ in the time−ordering of their nodes

These two are different:



These two are the same:



The number of labelled histories

The labelled history is essentally a list of the pairs of lineages that coalesce, in order. So the number of these is

$$\frac{n(n-1)}{2}\ \frac{(n-1)(n-2)}{2}\ \frac{(n-2)(n-3)}{2}\cdots\frac{2\times 1}{2}$$

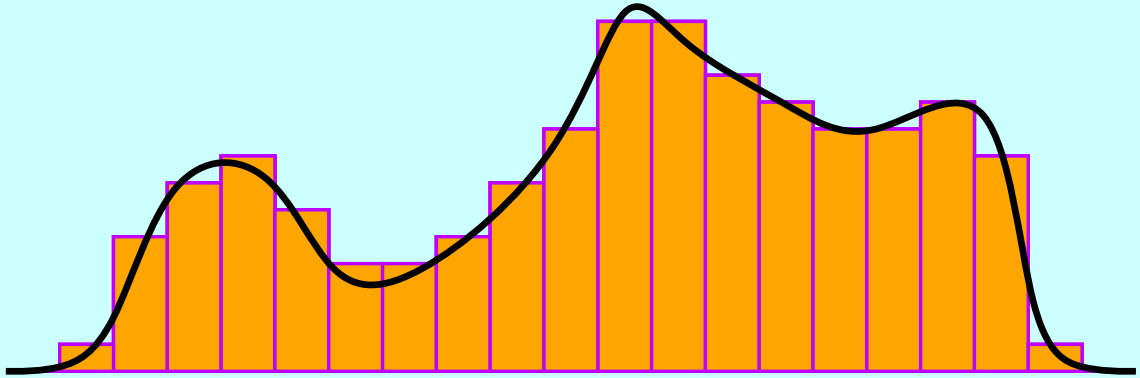
$$= \frac{n!(n-1)!}{2^{n-1}}$$

The number of these rises rapidly:

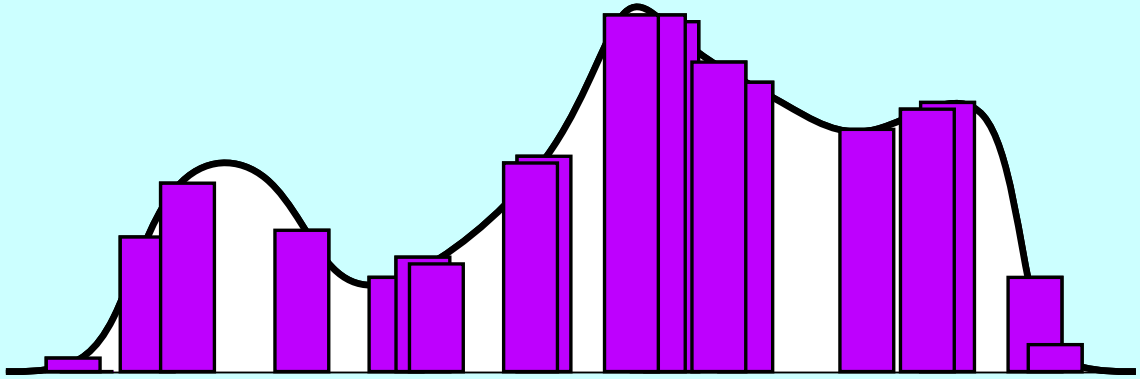
Tips	Labelled histories
2	1
3	3
4	18
5	180
6	2700
7	56,700
8	1,587,600
9	57,153,600
10	2,571,912,000

Monte Carlo integration

To get the area under a curve, we can either evaluate the function ($f(x)$) at a series of grid points and add up heights $\times$ widths:



or we can sample at random the same number of points, add up height $\times$ width:

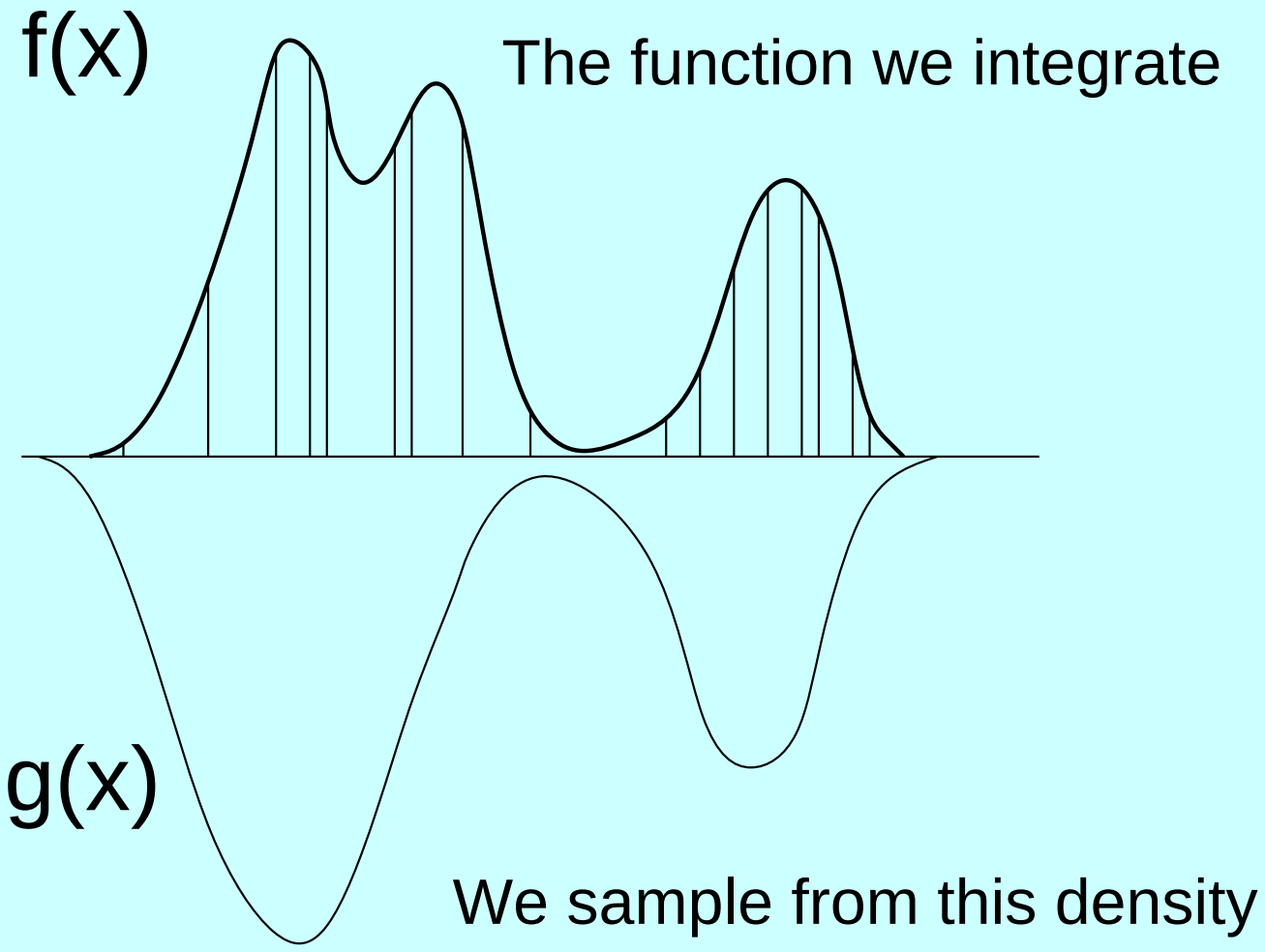


The importance sampling formula

Expectation of a function $h(x)$ over a distribution whose density function is $g(x)$:

$$E_g[h(x)] = \int_x h(x)g(x)\,dx$$

Importance Sampling



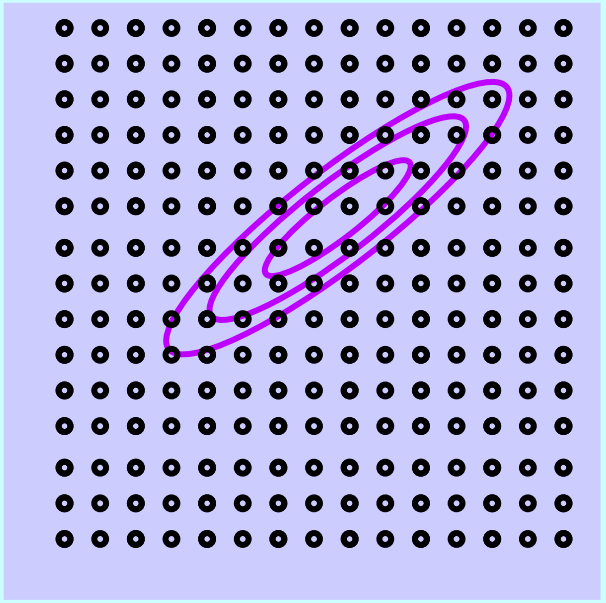
The integral can be computed as follows:

$$\begin{aligned}
\int f(x)\,dx &= \int \frac{f(x)}{g(x)}\, g(x)\,dx \\
&= E_g\left[\frac{f(x)}{g(x)}\right] \\
&\simeq \frac{1}{n}\sum_{i=1}^{n}\frac{f(x_i)}{g(x_i)}
\end{aligned}$$

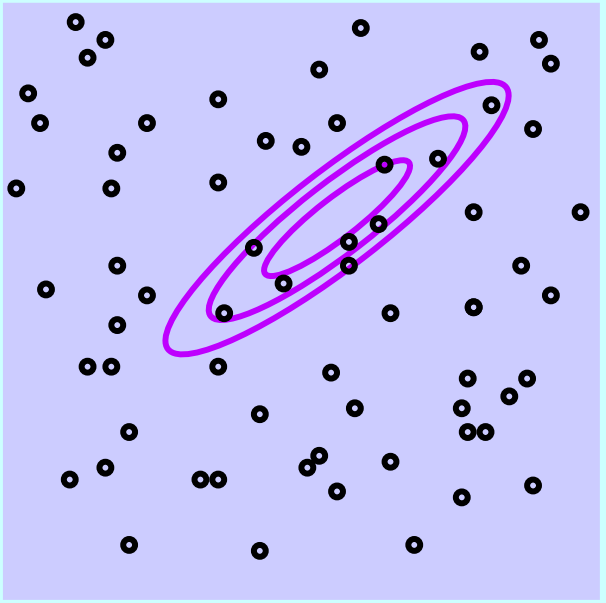
(where the sample points x_i are drawn from density $g(x)$)

In effect, each point sampled is taken to be a histogram bar whose width is less if it sampled from an area where $g(x)$ is larger, so that more samples get taken from that area.

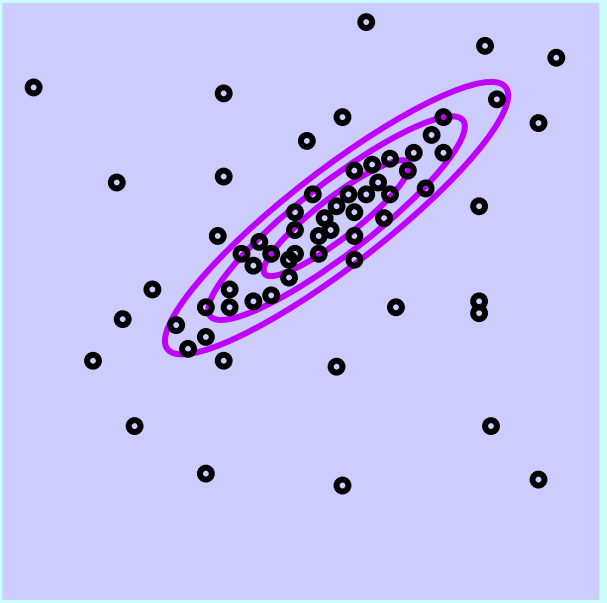
Accuracy is improved by importance sampling



Grid



Monte Carlo



Importance

Transition probabilities that achieve a given distribution

If we desire a particular equilbrium distribution π_i then one way to achieve it is to run a Markov chain that has transition probabilities that achieve *detailed balance*, so that for each pair of states the fraction of cases that move from i to j is the same as the fraction that move from j to i. If P_{ij} is the conditional probability of going from i to j then we achieve this if:

$$\pi_i \, P_{ij} \; = \; \pi_j \, P_{ji}$$

So if g_i is proportional to the desired distribution,

$$P_{ij}/P_{ji} = g_j/g_i$$

Any choice of P's that satisfies this is OK. To move around as fast as possible, suppose $g_j > g_i$. Then when j is proposed from i, accept it always. When i is proposed from j, accept it with probability g_i/g_j. So we use $P_{ij} = 1$ and $P_{ji} = g_i/g_j$.

MCMC: The Metropolis-Hastings method

To draw a sample $G_1, \ldots, G_n$ from a distribution proportional to a function $g(G)$:

(1) Draw a change in G from some "proposal distribution": $x \to y$

(2a) (Metropolis et. al., 1953):
Accept the change if a uniformly-distributed random number R satisfies

$$R < \frac{g(y)}{g(x)}$$

(2b) Hastings (*Biometrika*, 1970) corrected for biases toward some y's in the proposal distribution by using instead

$$R < \frac{\mathrm{Prob}(x|y)}{\mathrm{Prob}(y|x)} \frac{g(y)}{g(x)}$$

Repeat many times. If we do this long enough, and various niceness conditions hold, then $G_1, \ldots, G_m$ will be a sample from the right distribution.

An aside – the Gibbs Sampler

If the proposal distribution is proportional to the relative probabilities $g(y)$ of all states obtained by removing one coordinate z_i from $(z_1, z_2, \ldots, z_n)$ and then sampling proportional to the relative probabilities of all possible values that could be put there ...

Then the Hastings ratio for going from x to y is $g(x)/g(y)$, which leads to a perfect cancellation – you should always accept!

When you can do this it is generally a very fast method.

Computing coalescent likelihoods by MCMC

We want to compute $\int_G \mathrm{Prob}\,(G|\Theta)\mathrm{Prob}\,(D|G)dG$. We use an importance sampling density proportional to the interior of the integral at some trial value Θ_0 of the parameter. Then it is

$$g(G) = \frac{\mathrm{Prob}\,(G|\Theta_0)\mathrm{Prob}\,(D|G)\,dG}{\int_G \mathrm{Prob}\,(G|\Theta_0)\mathrm{Prob}\,(D|G)\,dG}$$

whose denominator is

$$L(\Theta_0) = \int_G \mathrm{Prob}\,(G\,|\,\Theta_0)\mathrm{Prob}\,(D\,|\,G)\,dG$$

The integral is:

$$L(\Theta) = \frac{1}{n}\sum_{i=1}^{n}\frac{f(G_i)}{g(G_i)}$$

$$= \frac{1}{n}\sum_{i=1}^{n}\frac{\text{Prob}\,(G|\Theta)\text{Prob}\,(D|G)}{\text{Prob}\,(G|\Theta_0)\text{Prob}\,(D|G)/L(\Theta_0)}$$

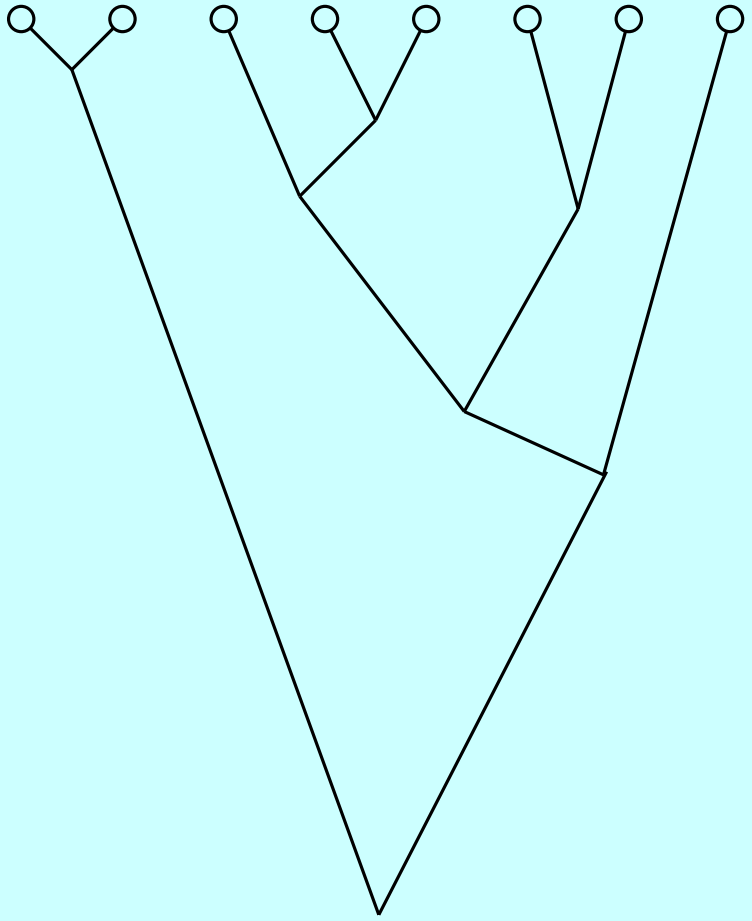
This leads to

$$\frac{L(\Theta)}{L(\Theta_0)} = \frac{1}{n}\sum_{i=1}^{n}\frac{f(G_i)}{g(G_i)} = \frac{1}{n}\sum_{i=1}^{n}\frac{\text{Prob}\,(G_i|\Theta)}{\text{Prob}\,(G_i|\Theta_0)}$$

Note that in the computation of the likelihood ratio, the terms for the probability of the data have disappeared! Where is the data? Its influence is on the importance sampling, where in tree space the trees are sampled from.

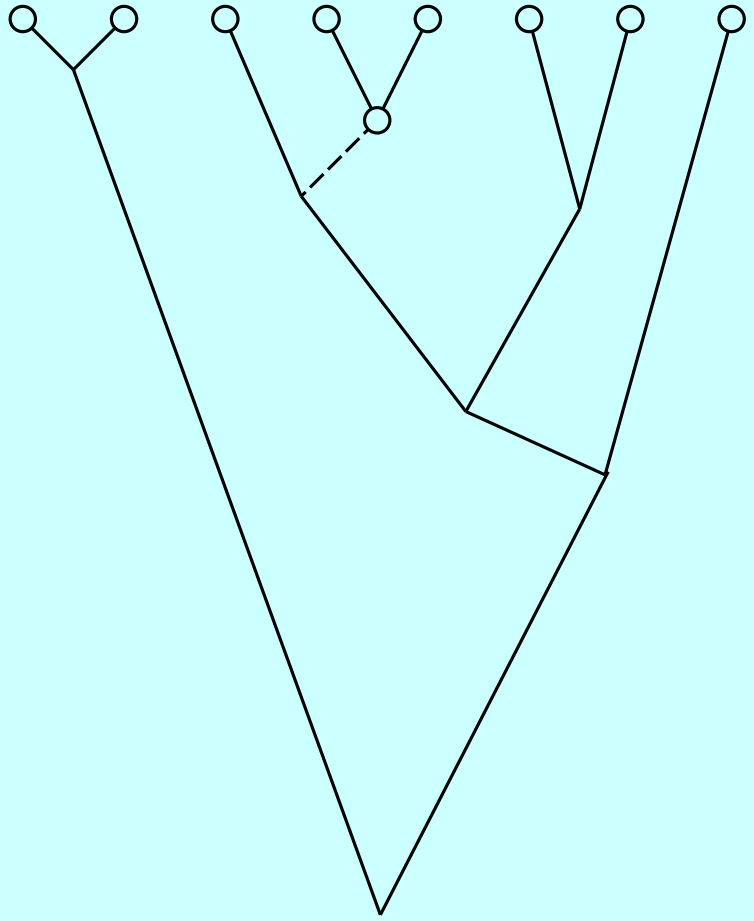
Tree rearrangements proposed:

A conditional coalescent rearrangement strategy



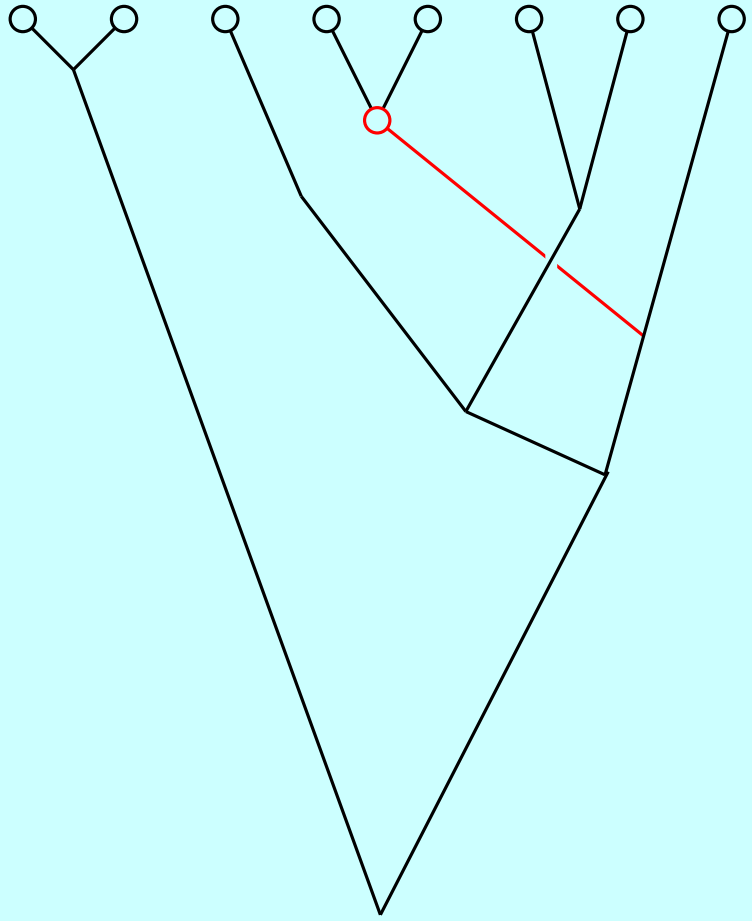
more ...

First pick a random node (interior or tip) and remove its subtree



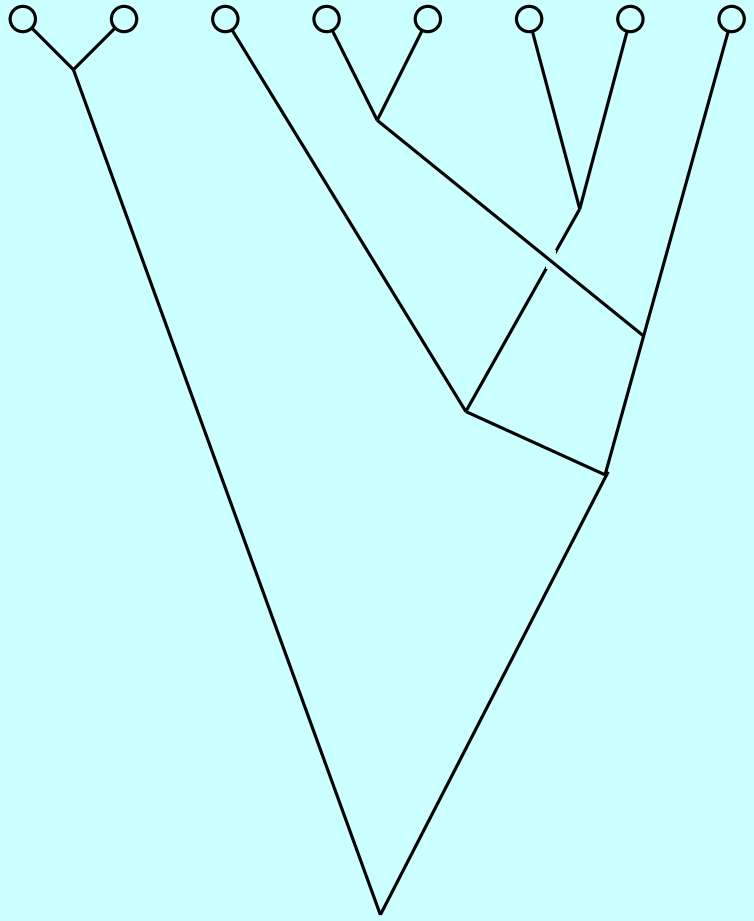
more ...

Then allow this node to re−coalesce with the tree



and finally we get:

The resulting tree proposed by this process

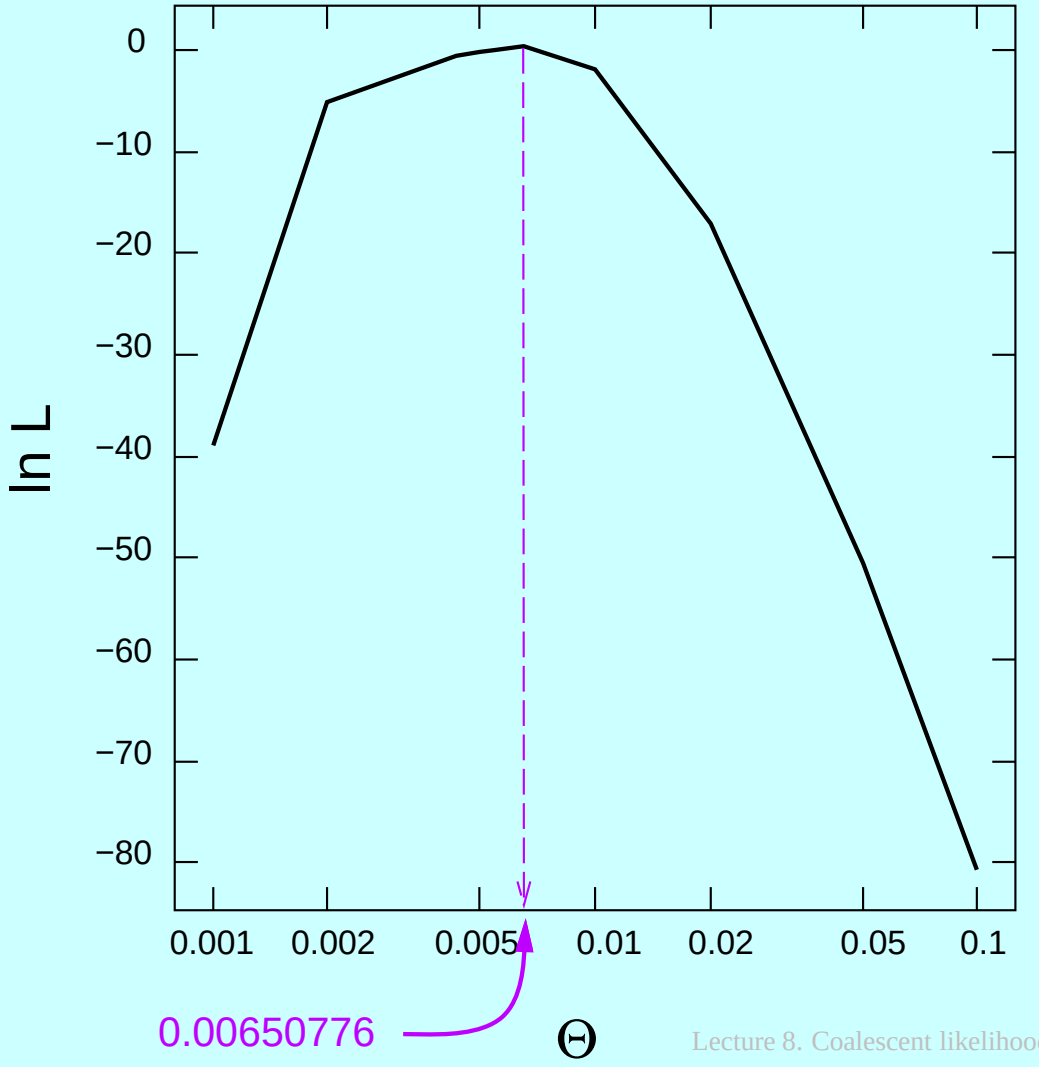


Being left out of the story:

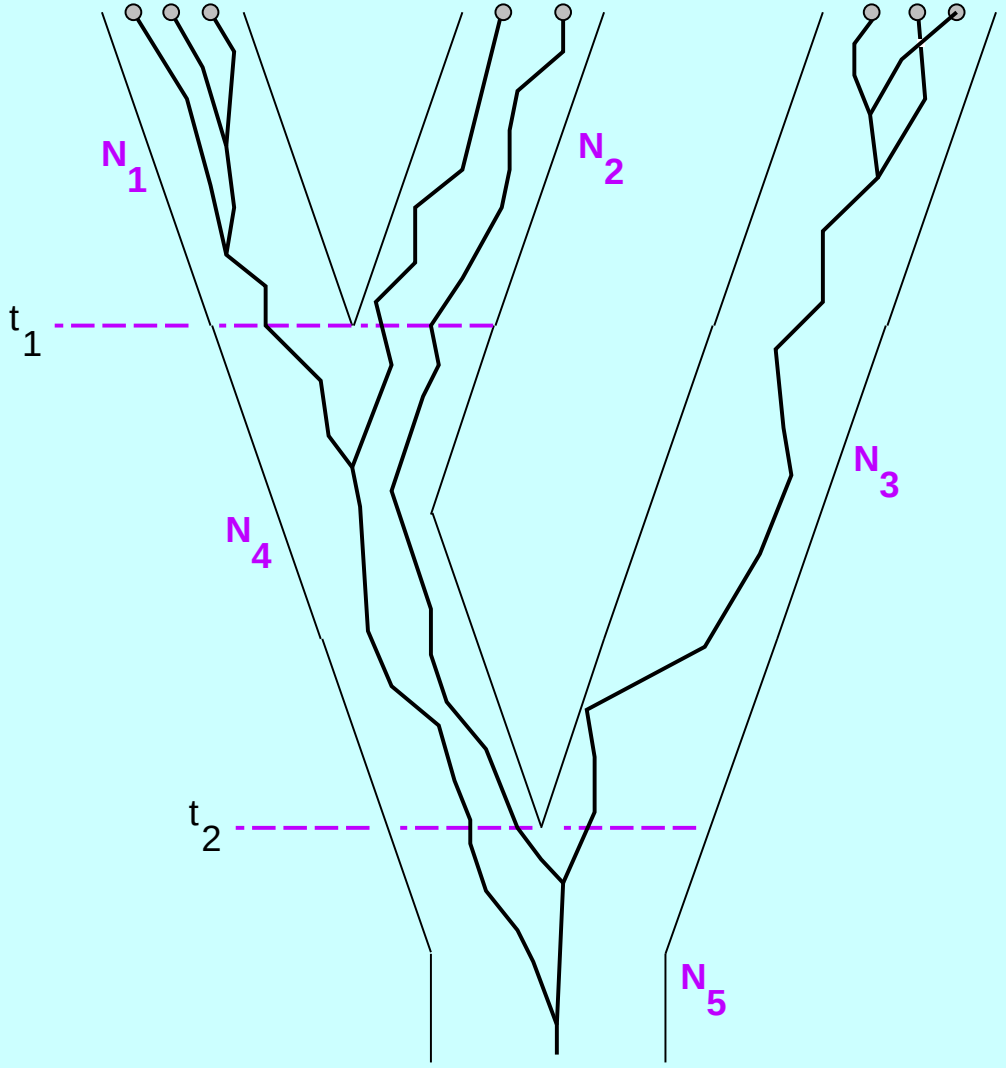
- We choose the rearrangements so that the proposal distribution is a “conditional coalescent".
- We do a Hastings correction given this.
- The end result is a perfect cancellation (which is pleasant rather than essential).
- This leaves us with the rule that we use $\mathrm{Prob}\left(\mathsf{D}\,|\,\mathsf{G}\right)$ as the only function in the Metropolizing.

One ends up with a curve that might look like this:

Results of analysing a data set with 50 sequences of 500 bases which was simulated with a true value of $\Theta = 0.01$

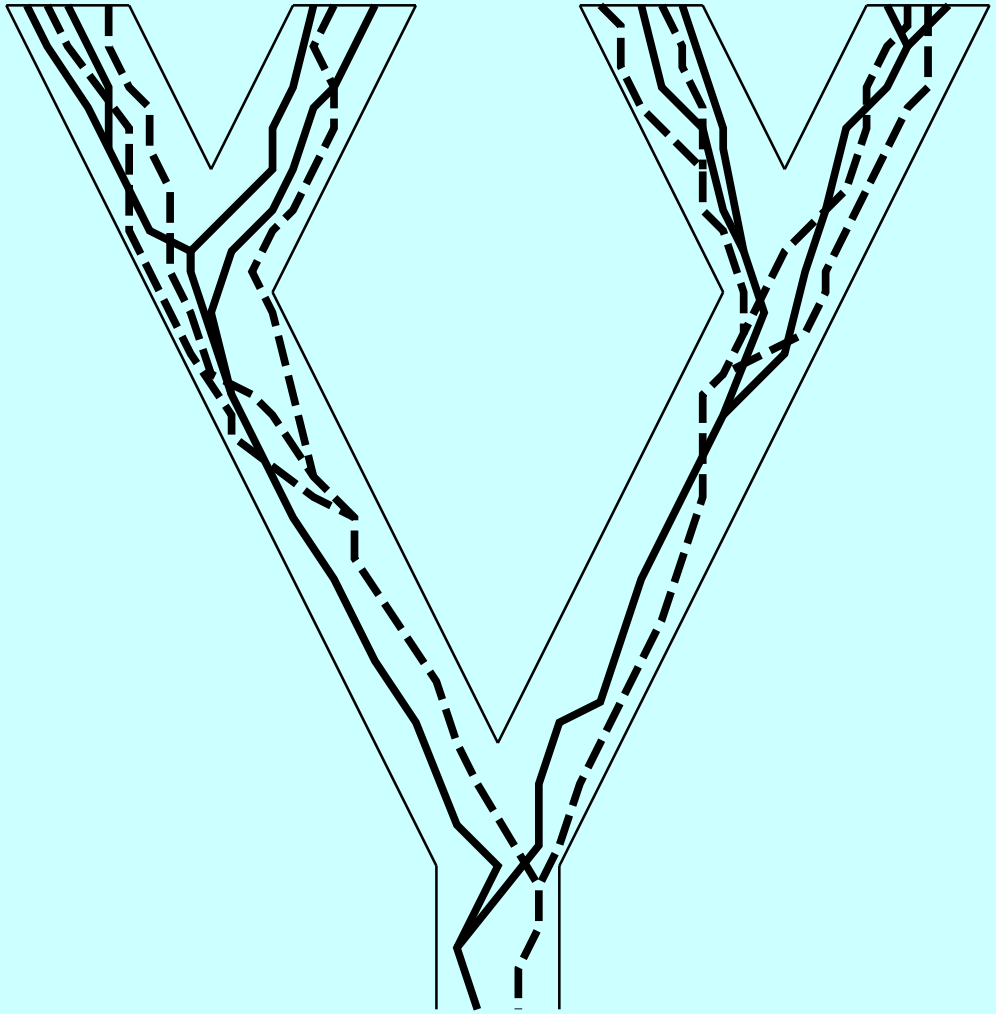


A coalescent and a species tree



Integrating over all coalescents

Evaluating fit of multiple loci to a species tree



Summing over all possible genealogies at each locus

$$L = \text{Prob}\,(\text{Data} \mid \text{Species tree})$$

$$= \prod_{\text{loci}} \sum_{\substack{\text{coalescent} \\ \text{trees}}} \text{Prob}\,(\text{coalescent } i \mid \text{Species tree}) \times \text{Prob}\,(\text{Data } i \mid \text{coalescent } i)$$

This involves integrating over all coalescent trees, separately at each locus (if the loci are not closely linked). If they are closely linked then one would jointly integrate them over a coalescent which has recombinations.

New genetic tools being deployed

Likelihood or Bayesian inference using sampling methods with coalescents

Mig = Migration, Rec = Recombination, Grow = Population growth, Split = Splittings, Bayes = Bayesian

Program Name	Mig?	Rec?	Grow?	Split?	Bayes?
LAMARC (Kuhner, Yamato et al.)	Y	Y	Y	n	y/n
BEAST (Drummond, Rambaut)	n	n	Y	Y	Y
Genetree (Griffiths and Bahlo)	Y	n	Y	n	n
Migrate (Beerli)	Y	n	n	n	Y
IM, Ima (Nielsen & Hey)	Y	n	y/n	Y	Y
Batwing (Wilson and Balding)	n	n	n	Y	Y

References

Edwards, A. W. F. 1970. Estimation of the branch points of a branching diffusion process. *Journal of the Royal Statistical Society B* **32:** 155-174. [Labelled histories]

Metropolis, N., A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller and E. Teller. 1953. Equation of state calculations by fast computing machines. *Journal of Chemical Physics* **21:** 1087-1092. [The Metropolis algorithm]

Hastings, W. K. 1970. Monte Carlo sampling methods using Markov chains and their applications. *Biometrika* **57:** 97-109. [The Hastings correction]

Geman, S. and D. Geman. 1984. Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **6:** 721-741. [The original paper on the Gibbs sampler]

Griffiths, R. C. 1989. Genealogical tree probabilities in the infinitely-many-site model. *Journal of Mathematical Biology* **27:** 667-680. [Summing up over event histories]

References

Felsenstein, J. 1992. Estimating effective population size from samples of sequences: inefficiency of pairwise and segregating sites as compared to phylogenetic estimates. *Genetical Research* **59:** 139-147. [Suggests using the coalescents]

Griffiths, R. C. and S. Tavaré. 1994a. Sampling theory for neutral alleles in a varying environment. *Philosophical Transactions of the Royal Socety of London, Series B (Biological Sciences)* **344:** 403-10. [Griffiths-Tavaré sampling method]

Griffiths, R. C. and S. Tavaré. 1994b. Ancestral inference in population genetics. *Statistical Science* **9:** 307-319. [Griffiths-Tavaré sampling method]

Kuhner, M. K., J. Yamato, and J. Felsenstein. 1995. Effective population size and mutation rate from sequence data using Metropolis-Hastings sampling. *Genetics* **140:** 1421-1430. [Our MCMC coalescent likelihood method]

References

Kuhner, M. K., J. Yamato and J. Felsenstein. 1998. Maximum likelihood estimation of population growth rates based on the coalescent *Genetics* **149:** 429-434. [Our approach to growing populations, and describes a bias in estimation]

Griffiths, R. C. and P. Marjoram. 1996. Ancestral inferences from samples of DNA sequences with recombination. *Journal of Computational Biology* **3:** 479-502. [Coalescent likelihoods with recombination]

Kuhner, M. K., J. Yamato, and J. Felsenstein. 2000. Maximum likelihood estimation of recombination rates from population data. *Genetics* **156:** 1393-1401. [Our approach to coalescent likelihoods with recombination]

Beerli, P. B. and J. Felsenstein. 1999. Maximum-likelihood estimation of migration rates and effective population numbers in two populations using a coalescent approach. *Genetics* **152:** 763-773. [Our approach to migration estimation]

References

Griffiths, R. C. and S. Tavaré. 1997. Computational methods for the coalescent. pp. 165-182 in *Progress in Population Genetics and Human Evolution*, ed. P. Donnelly and S. Tavaré. IMA Volumes on Mathematics and Its Applications, volume 87. Springer, New York. [Review of their approach]

Felsenstein, J., M. K. Kuhner, J. Yamato, and P. Beerli. 1999. Likelihoods on coalescents: a Monte Carlo sampling approach to inferring parameters from population samples of molecular data. pp. 163-185 in *Statistics in Molecular Biology and Genetics*, ed. F. Seillier-Moiseiwitsch. IMS Lecture Notes-Monograph Series, volume 33. Institute of Mathematical Statistics and American Mathematical Society, Hayward, California. [Overview of my lab's methods. Correct discussion of their relation to Griffiths and Tavaré's methods.]

Stephens, M. and P. Donnelly. 2000. Inference in molecular population genetics. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **62:** 605-635. [Major speedup of the Griffiths-Tavaré approach]

References

Nielsen, R. 1997. Maximum likelihood estimation of population divergence times and population phylogenies under the infinite sites model. *Theoretical Population Biology* **53:** 143-151. [The first coalescent likelihood paper with more than one species]

Hein, J., M. Schierup, and C. Wiuf. 2005, *Gene Genealogies, Variation and Evolution: A Primer in Coalescent Theory.* Oxford University Press, Oxford. [One of two books so far on coalescents. Light on estimation issues]

Wakeley, J. 2008. *Coalescent Theory.* Roberts and Co., Greenwood Village, Colorado. [One of two books so far on coalescents. Light on estimation issues.]

Felsenstein, J. 2004. *Inferring Phylogenies.* Sinauer Associates, Sunderland, Massachusetts. [Especially chapter 27 which covers MCMC likelihood approaches (but explanation of logic of Griffiths/Tavaré method is wrong)]