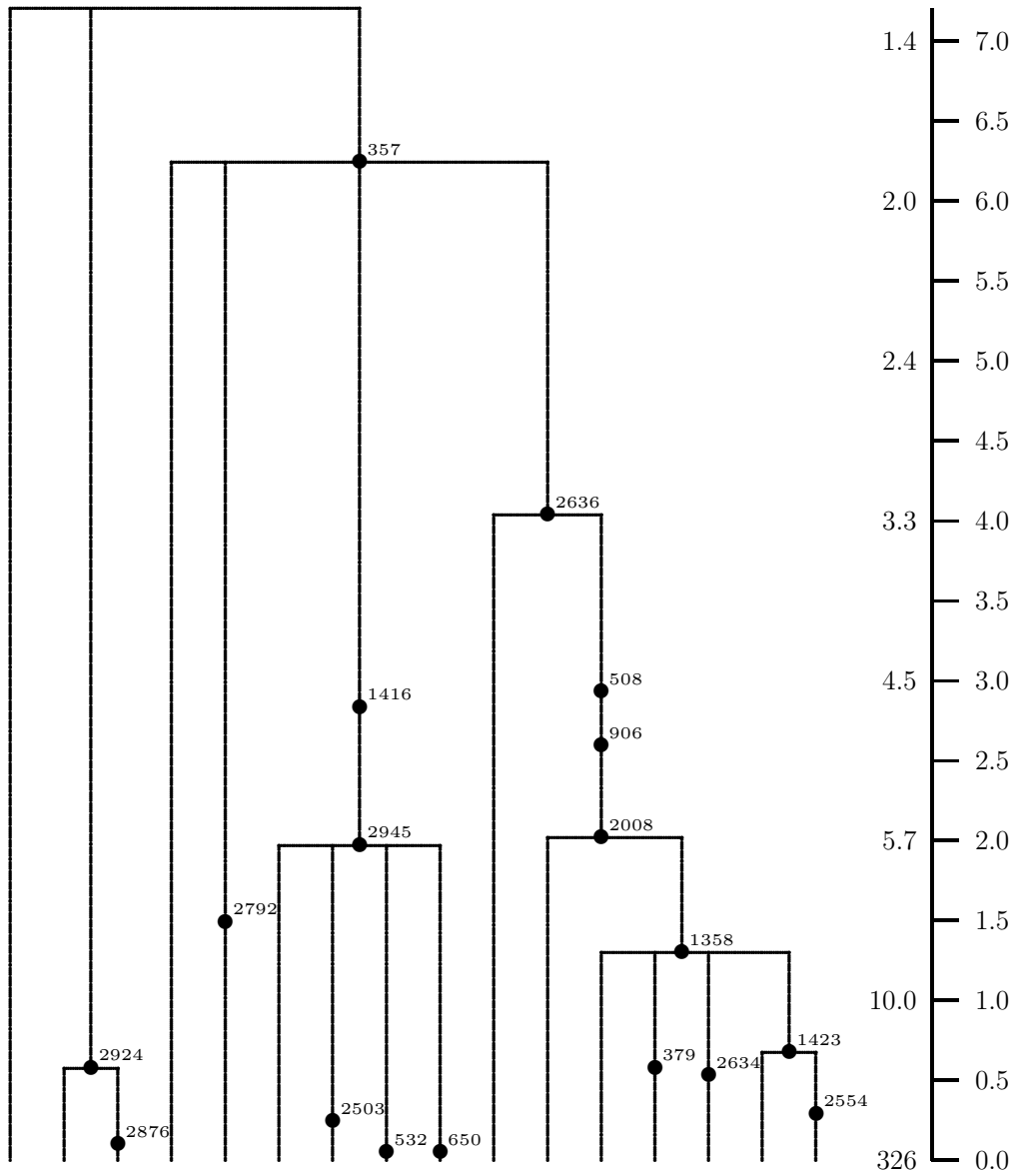# Stochastic models
in Mathematical Genetics

Hilary Term 06

Professor R. C. Griffiths
Department of Statistics
University of Oxford

World tree from β-globin data.
Time units are in 100,000 years.

	B2	B4	B11	B3	B1	A1	A3	A2	A4	B9	C3	C2	C1	C7	D2	D1
World	18	3	1	9	79	104	8	1	1	2	10	9	48	19	1	13
Pygmies	4	·	·	1	6	9	1	·	·	1	·	·	·	·	·	·
Gambia	6	3	·	2	5	8	2	·	·	·	·	·	1	·	1	·
Kenya	8	·	1	6	9	12	5	·	·	·	·	·	·	·	·	1
Mongolia	·	·	·	·	3	3	·	·	1	·	·	2	4	6	·	3
Amerind	·	·	·	·	2	15	·	·	·	·	·	6	22	·	·	1
PNG	·	·	·	·	12	1	·	·	·	·	7	·	·	4	·	·
Sumatra	·	·	·	·	10	8	·	·	·	·	·	1	14	6	·	·
UK	·	·	·	·	16	23	·	·	·	1	·	·	·	2	·	4
Vanuatu	·	·	·	·	16	25	·	1	·	·	3	·	7	1	·	4

This gene tree is constructed from 326 β-globin DNA sequences.

Vertices represent mutations at different positions along the DNA sequences.

B2, B4, ... are labels of sequence types (haplotypes), and the numbers are split into subpopulation numbers where sequences were collected. The time scale on the right is in units of 100,000 years. The inferred time to the most recent common ancestor of the sample of sequences is over 700,000 years. Expected numbers of ancestors are to the left of the axis on the right.

This is an example of an evolutionary gene tree. Mathematical tools used in its construction are graph theory, stochastic processes, statistical inference and computationally intensive methods. DNA data is from research conducted at the Institute of Molecular Medicine, University of Oxford. The reference for the paper with the gene tree is:

Harding R. M., Fullerton S. M., Griffiths R. C., Bond J., Cox M. J., Schneider J. A., Moulin D., and Clegg J. B. (1997).
Archaic African *and* Asian lineages in the genetic ancestry of modern humans.
American Journal of Human Genetics, **60**, 772–789.

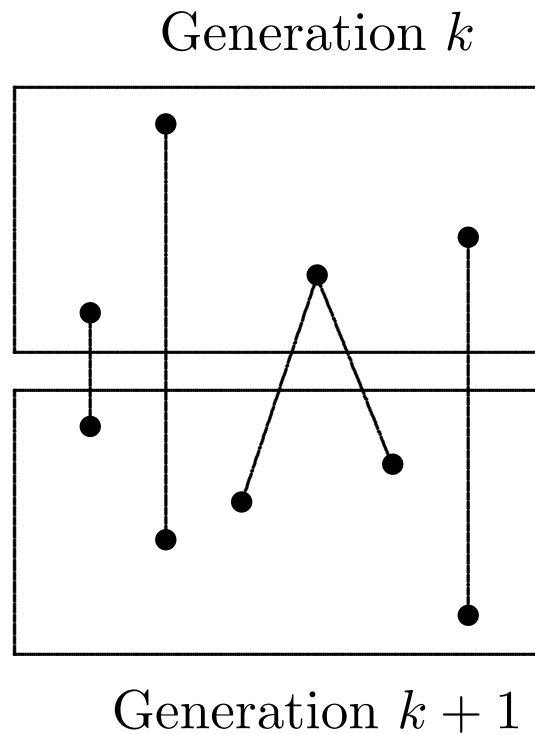# Evolutionary models

Wright-Fisher model

A population of M genes.
Discrete generations.
Reproductive mechanism:
Generation $k+1$ is formed from generation k by choosing M genes at random with replacement. Some genes in generation k may have no off-spring, while others have multiple offspring.

Generation k

Generation $k+1$

Problem. What is the probability p_{ij} that i genes from generation $k+1$ have j parents in generation k?

Answer. This is the probability that when i balls are placed at random in M boxes, exactly j boxes are non-empty. Here parents are identified with `boxes` and offspring as `balls`.
The distribution (for i fixed) is

$$p_{ij} = \mathcal{S}(i,j)\frac{M_{[j]}}{M^i}, \ j \leq i,$$

where

$$M_{[j]} = M(M-1)\cdots(M-j+1)$$

and $\{\mathcal{S}(i,j)\}$ are Stirling numbers of the second kind, defined by coefficients in the expansion

$$x^i = \sum_{j=1}^{i}\mathcal{S}(i,j)x_{[j]}.$$

A formula is

$$\mathcal{S}(i,j) = \frac{1}{j!}\sum_{k=0}^{j}(-1)^{j-k}\binom{j}{k}k^i.$$

$\mathcal{S}(i,j)$ is the number of ways of partitioning a set of i elements into j non-empty subsets.
exercise: $\sum_{j=1}^{M}p_{ij} = 1$.

The number of ancestors $\{\xi(\tau), \tau = 0, 1, \ldots\}$ of a sample of $\xi(0) = i$ genes at generation τ back in time is a homogeneous Markov chain with transition matrix P. This assumes that the population can be extended back in time forever, if not and the founding generation is τ_0 generations back, the process stops with a random number of ancestors $\xi(\tau_0)$ in the founders.
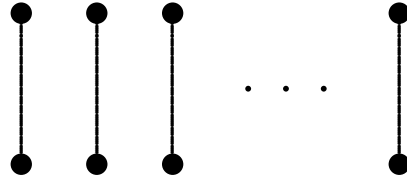
The population size M is usually assumed to be large, then with high probability the number of ancestors of a sample of j genes one generation back is either j or $j - 1$.

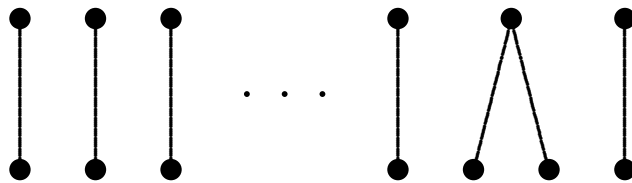$$p_{j,j} = \frac{M}{M} \cdot \frac{M-1}{M} \cdots \frac{M-j+1}{M}$$

$$p_{j,j-1} = \binom{j}{2} \frac{1}{M}$$

$$\cdot \frac{M-1}{M} \cdots \frac{M-1-(j-2)+1}{M}.$$

The formula for $p_{j,j-1}$.

Distinct parents

Two with the same parent

There are $\binom{j}{2}$ unordered pairs of genes from the j to choose as a pair with the same parent. The probability is $1/M$ that the second gene has the same parent as the first. The other $j-2$ have distinct parents with probability

$$\frac{M-1}{M} \cdots \frac{M-1-(j-2)+1}{M}.$$

As $M \to \infty$,

$$p_{j,j} = \left(1 - \frac{1}{M}\right)\left(1 - \frac{2}{M}\right)\cdots\left(1 - \frac{j-1}{M}\right)$$

$$\approx 1 - \sum_{\ell=1}^{j-1} \frac{\ell}{M}$$

$$\approx 1 - \frac{j(j-1)}{2M}$$

$$p_{j,j-1} \approx \frac{j(j-1)}{2M}$$

$$p_{j,i} = O\left(M^{-(j-i)}\right)$$

Terms omitted are of order M^{-2}.

The coalescent.

If time is measured in units of M generations, and $M \to \infty$, then the ancestral tree in the Wright-Fisher model converges to a *coalescent tree.*

Two ancestral lineages coalesce when they have a common ancestor forming a vertex in the tree. The coalescent process is quite famous in Mathematical Genetics and has an Oxford connection of being developed by Sir John Kingman who was a Professor at Oxford.

Kingman, J. F. C. (1982). The coalescent. *Stochastic Processes and their Applications* **13**, 235–248.

Coalescent tree.
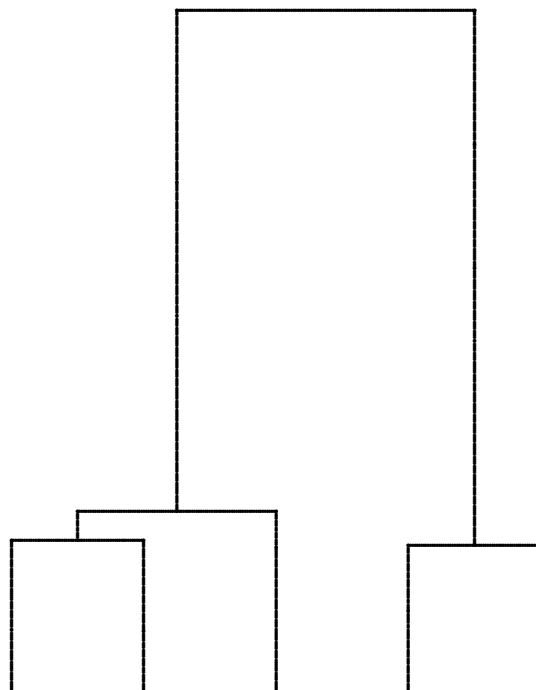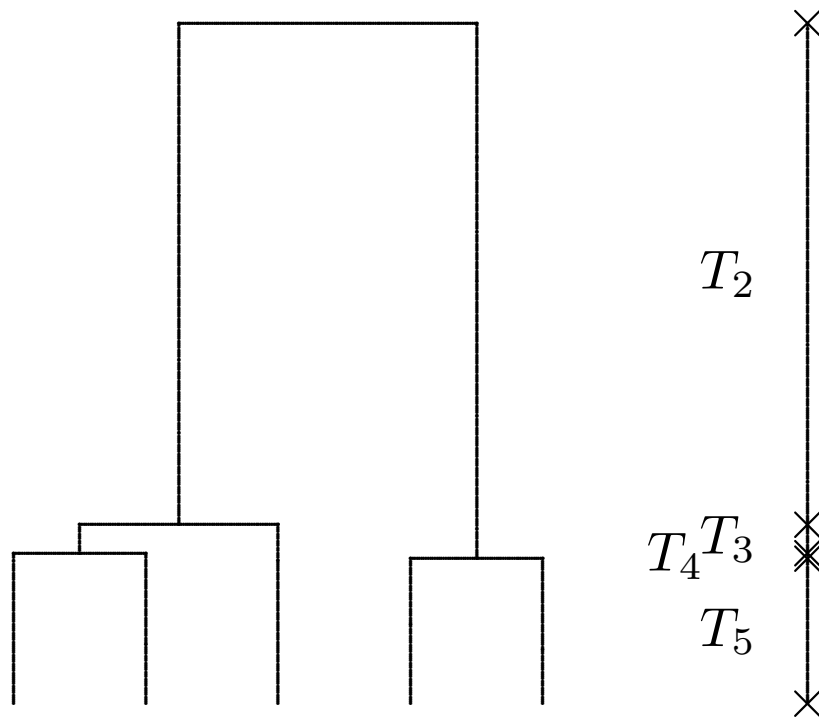
Coalescent tree of $n = 5$ genes.
$T_n, T_{n-1}, \ldots, T_2$ are the times while
$n, n-1, \ldots, 2$ ancestors of the sample.
$\{T_j; j = n, \ldots, 2\}$ are independent exponential
random variables with

$$E(T_j) = \frac{2}{j(j-1)}.$$

The probability density function of T_j is

$$f_j(t) = \binom{j}{2} \exp\left(- \binom{j}{2} t \right), \ t > 0.$$

Wright-Fisher limit

Let $\tau_n, \tau_{n-1}, \ldots \tau_2$ be the times spent while $n, n-1, \ldots, 2$ ancestor genes in the Wright-Fisher model. Distribution of τ_2.

$$p_{22} = 1 - \frac{1}{M}, \; p_{21} = \frac{1}{M},$$

$$P(\tau_2 = k) = p_{22}^{k-1} p_{21}$$
$$= \frac{1}{M}\left(1 - \frac{1}{M}\right)^{k-1}, \; k = 1, 2, \ldots$$

τ_2 has a geometric distribution (starting at 1), with mean $E(\tau_2) = M$. Measuring time in units of M, so that $T_2^{(M)} = \tau_2/M$,

$$P(T_2^{(M)} \le t) = P(\tau_2 \le [Mt])$$
$$= 1 - \left(1 - \frac{1}{M}\right)^{[Mt]}$$
$$\to 1 - e^{-t}, \text{ as } M \to \infty.$$

Thus $T_2^{(M)}$ has a limit exponential distribution.

Let the scaled waiting time while r ancestors be $T_r^{(M)} = \tau_r/M$.

Then

$$P(T_r^{(M)} \leq t) = 1 - p_{rr}^{([Mt])}$$

$$P(T_r^{(M)} \leq t) = 1 - p_{rr}^{[Mt]}$$

$$\approx 1 - \left(1 - \frac{\binom{r}{2}}{M}\right)^{[Mt]}$$

$$\rightarrow 1 - \exp\left(-\binom{r}{2}t\right).$$

$T_r^{(M)}$ has a limit exponential distribution with mean $\frac{2}{r(r-1)}$.

It follows that in the coalescent limit the times $T_n, \ldots, T_2$ while $n, \ldots, 2$ ancestors are independent exponential random variables with T_j having a probability density function

$$\binom{j}{2} \exp\left\{-\binom{j}{2}t\right\}, \ t > 0.$$

The time to the most recent common ancestor (TMRCA) of n genes is

$$W_n = T_n + T_{n-1} + \cdots + T_2.$$

The mean time to the MRCA is

$$E(W_n) = \sum_{j=2}^{n} \frac{2}{j(j-1)}$$

$$= 2 \sum_{j=2}^{n} \left[\frac{1}{j-1} - \frac{1}{j} \right]$$

$$= 2 \left(1 - \frac{1}{n} \right).$$

Coalescence is so fast that it is possible to have a coalescent tree with an infinite number of genes (thought of as the whole population).
The TMRCA is

$$W_\infty = \sum_{j=2}^{\infty} T_j,$$

and $E(W_\infty) = 2$.
In units of generations $E(W_\infty) = 2M$ generations, so if the generation time is G years,

$$E(W_\infty) = 2MG \text{ years.}$$

The number of ancestors at time t back.

Let $A_n(t)$ be the number of ancestors of a sample of n genes at time t back from the present. Then
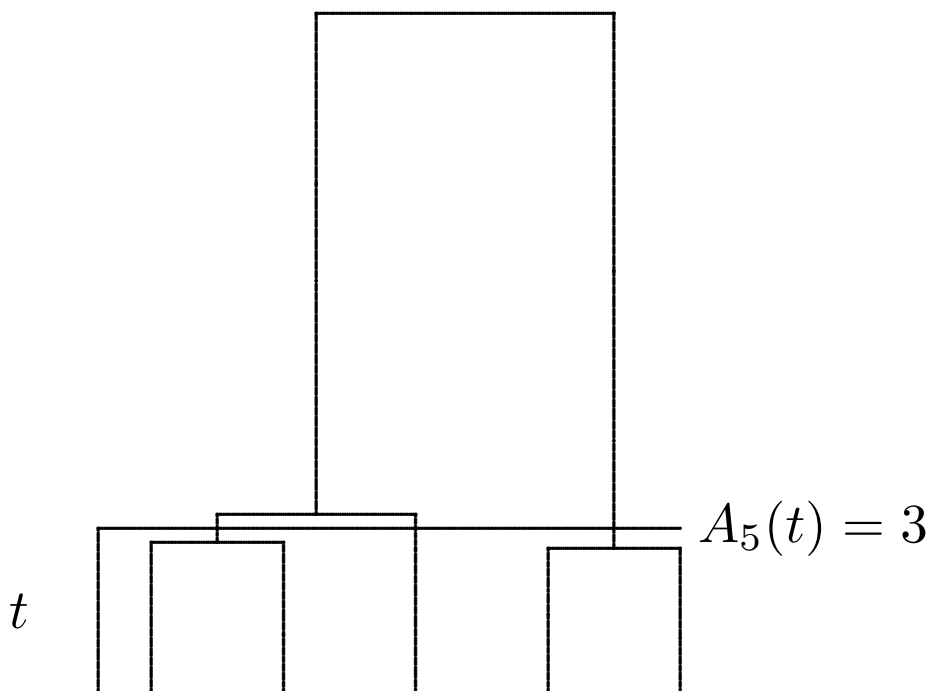
$$\{A_n(t) \leq k\} \equiv \{T_n + \cdots + T_k \leq t.\}$$

$\{A_n(t), t \geq 0\}$ is a death process in the sense of a birth and death process in probability theory with zero birth rates and death rates

$$\mu_k = \binom{k}{2}, \ k = n, n-1, \ldots, 2.$$

In the coalescent tree $A_n(t)$ is the number of edges in a cross section of the tree at time t back.

$A_5(t) = 3$

t

Denote the ancestor transition functions as

$$g_{ij}(t) = P(A(t) = j \mid A(0) = i).$$

It is possible to show that for $2 \leq j \leq i$,

$$g_{ij}(t) = \sum_{k=j}^{i} \rho_k(t) \frac{(2k-1)(-1)^{k-j} j_{(k-1)} i_{[k]}}{j!(k-j)! i_{(k)}},$$

where $\rho_k(t) = \exp\left(-\binom{k}{2}t\right)$, and

$$g_{i1}(t) = 1 - \sum_{k=2}^{i} \rho_k(t) \frac{(2k-1)(-1)^k i_{[k]}}{i_{(k)}}.$$

In these equations

$$\alpha_{[k]} = \alpha(\alpha - 1) \ldots (\alpha - k + 1)$$
$$\alpha_{(k)} = \alpha(\alpha + 1) \ldots (\alpha + k - 1)$$

The mean is

$$E\left(A_i(t)\right) = \sum_{\ell=1}^{i} \rho_\ell(t)(2\ell - 1) \frac{i_{[\ell]}}{i_{(\ell)}}.$$

In the whole population with $i = \infty$ the formulae still hold, with $i_{[\ell]}/i_{(\ell)} = 1$.
A reference is
Tavaré, S. (1984) Line-of-descent and genealogical processes and their applications in population genetics models.
Theoretical Population Biology, **26**, 119–164.

Problem. At a time instant suppose that $A_i(t) = k$. Let $Z_1, \ldots Z_k$ be the number of genes subtended by the k edges $Z_1 + Z_2 + \cdots + Z_k = i$. What is the probability distribution of $Z_1, \ldots, Z_k$? We suppose the k edges are labelled in some way $1, \ldots, k$.
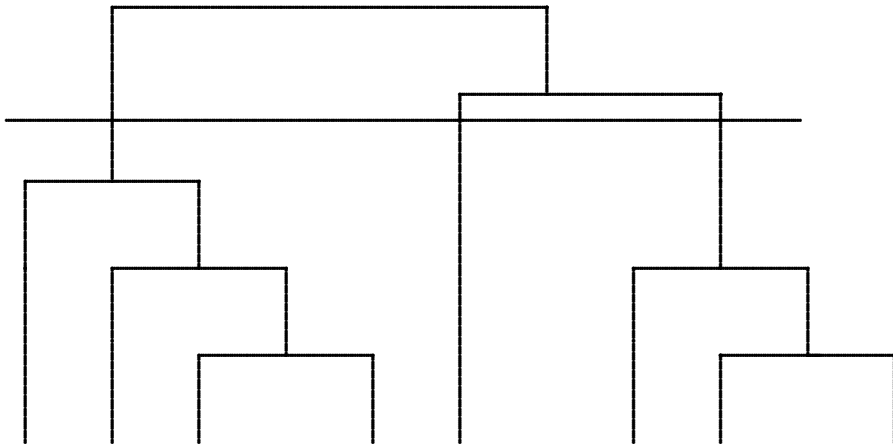
In the picture the three edges $(k = 3)$ cut by the horizontal line are such that there are $i = 8$ sample genes,

$$z_1 = 4, z_2 = 1, z_3 = 3.$$

Answer. $Z_1, \ldots, Z_k$ is uniform on partitions of i
$z_1 + z_2 + \cdots + z_k = i$, with

$$P(z_1, \ldots, z_k) = \binom{i-1}{k-1}^{-1}.$$

The proof is by induction on i. It is true for
$i = 2, k = 2$. Suppose it is true as an induction
hypothesis up to $i - 1, k = 2, \ldots, i - 2$.
If $k = i$ we are done. If $k < i$, then consider the
$i - 1$ edges before the last vertex.

$$P(z_1, \ldots, z_k) = \sum_{\ell=1}^{k} \frac{z_\ell - 1}{i - 1}$$
$$\cdot P(z_1, \ldots, z_{\ell-1}, z_\ell - 1, z_{\ell+1}, \ldots, z_k).$$

Coalescence is to an edge subtended by edge ℓ,
and the configuration when $i - 1$ lines is (for
$z_\ell > 1$) $z_1 + \cdots + z_{\ell-1} + z_\ell - 1 + \cdots + z_k = i - 1$.
The probability that edge ℓ branches is $(Z_\ell - 1)/(i - 1)$.
By the induction hypothesis the sum is

$$P(z_1, \ldots, z_k) = \sum_{\ell=1}^{k} \frac{z_\ell - 1}{i - 1} \binom{i-2}{k-1}^{-1}$$
$$= \frac{i - k}{i - 1} \binom{i-2}{k-1}^{-1}$$
$$= \binom{i-1}{k-1}^{-1}.$$

The induction proof is complete.

Problem. Take a single edge while k ancestors. What is the probability distribution of the number of individuals Z subtended in the sample of i below?

Answer. In the notation of the last problem suppose without loss of generality that $Z = Z_1$, and $Z' = Z_2 + \ldots + Z_k$. Then

$$Z + Z' = i$$

and for fixed $Z = z$ there are $\binom{i-z-1}{k-2}$ partitions of $z_2 + \cdots + z_k = z' = i - z$. Therefore

$$P(Z = z) = \frac{\binom{i-z-1}{k-2}}{\binom{i-1}{k-1}}.$$

Important note. This result holds for a general binary coalescent tree with an exchangeable coalescent structure. It doesn't depend on times $\{T_j\}$.

An urn model approach.

Identify edges in the tree with coloured balls in an urn. Let the k edges in a cross section be represented by k different coloured balls.

When the tree branches, in the direction of the root to the leaves, this is identified with adding a ball of the same colour as the parent edge to the urn.

The urn model is thus the following.

Begin with k different coloured balls in an urn. A ball is drawn out and replaced with one of the same colour. This is repeated until there are i balls in the urn.

A classical result in probability theory is that the numbers of different coloured balls $Z_1, \ldots, Z_k$ is uniform on the $\binom{i-1}{k-1}$ partitions of i into k parts. Taking a limit where $i \to \infty$, so the coalescent tree has an infinite number of leaves, the limit density of the relative proportions $X_1, \ldots, X_k$ of the k colours has a uniform distribution on the set $(x_1, \ldots, x_k)$ such that $0 < x_i < 1, i = 1, \ldots, k$ and $x_1 + \cdots + x_k = 1$.

If leaves in the coalescent tree are labelled $1, 2, \ldots, n,$ then the probability that edges in the tree while k edges have families of leaves $C_1, C_2, \ldots, C_k$ is

$$\frac{k!(k-1)!(n-k)!}{n!(n-1)!} \lambda_1! \ldots \lambda_k!,$$

where $\lambda_1, \ldots, \lambda_k$ are the sizes of $C_1, \ldots, C_k$. In this formulation $C_1, \ldots, C_k$ is a partition of $\{1, 2, \ldots, n\}$. The formula can be expressed as

$$\binom{n-1}{k-1}^{-1} k! \frac{\lambda_1! \ldots \lambda_k!}{n!}$$

Note that
$$\frac{\lambda_1! \ldots \lambda_k!}{n!}$$

is the probability of the (ordered) partition, given numbers $\lambda_1 + \lambda_2 + \cdots + \lambda_k = n$ and the term $k!$ comes from unordering $C_1, \ldots, C_k$.
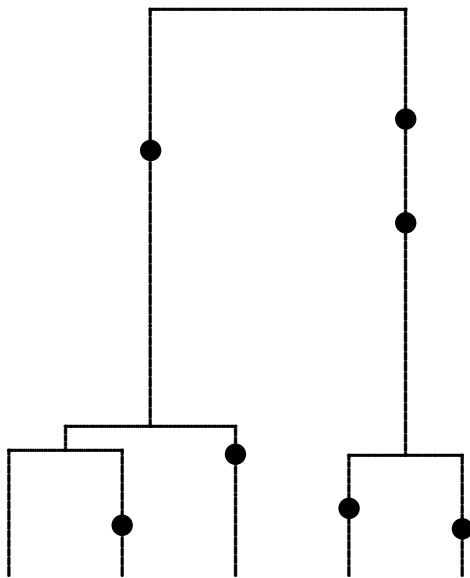
Mutations on genes.

Wright-Fisher model.

Mutations occur with probability u on offspring between generations. The total expected number of mutations per generation is thus Mu.

In the coalescent process the mutation rate u is scaled such that $\theta = 2Mu$.

Coalescent tree.

Mutations occur at a rate of $\theta/2$ on the edges of a coalescent tree in the coalescent time scale according to a Poisson process, given the edge lengths of the tree.

The number of mutations on a coalescent tree

Let $T_n, T_{n-1}, \ldots, T_2$ be times while $n, n-1, \ldots, 2$ ancestors of a sample of n genes. The random variables $\{T_j\}$ are independent and T_j has an exponential distribution with density

$$f_j(t) = \binom{j}{2} \exp\left(-\binom{j}{2}t\right), \ t > 0.$$

The total edge lengths in the tree are

$$nT_n + (n-1)T_{n-1} + \cdots + 2T_2,$$

so the number of mutations M on the coalescent tree has a probabilily generating function

$$P_n(z) = E \exp\left\{(z-1)\frac{\theta}{2}\left(nT_n + \cdots 2T_2\right)\right\}.$$

This is the *pgf* of a Poisson random variable with a random mean.

$$P_n(z) = \prod_{j=2}^{n} E\left(e^{(z-1)\frac{\theta}{2}jT_j}\right)$$

$$= \prod_{j=2}^{n} \left(1 - (z-1) \cdot \frac{j\theta/2}{j(j-1)/2}\right)^{-1}$$

$$= \prod_{j=2}^{n} \left(1 - \frac{(z-1)\theta}{(j-1)}\right)^{-1}$$

$$= \prod_{j=1}^{n-1} \left(1 - \frac{(z-1)\theta}{j}\right)^{-1}$$

If M_k is the number of mutations while k ancestors, then M_k has a geometric distribution

$$P(M_k = \ell) = \left(\frac{\theta}{k-1+\theta}\right)^\ell \frac{k-1}{k-1+\theta},$$
$$\ell = 0, 1, \ldots.$$

and $M = M_n + M_{n-1} + \ldots + M_2$.
The *pgf* of M_k is

$$Q_k(z) = \left(1 - \frac{(z-1)\theta}{(k-1)}\right)^{-1}.$$

$$Q'_k(1) = \frac{\theta}{k-1}, \; Q''_k(1) = 2\left(\frac{\theta}{k-1}\right)^2.$$

$$E(M_k) = \frac{\theta}{k-1}, \; \text{var}(M_k) = \left(\frac{\theta}{k-1}\right)^2 + \frac{\theta}{k-1}.$$

The expected number of mutations on the coalescent tree is

$$E(M) = \theta \sum_{j=1}^{n-1} \frac{1}{j},$$

and the variance is

$$\text{var}(M) = \sum_{j=1}^{n-1} \left\{\left(\frac{\theta}{j}\right)^2 + \frac{\theta}{j}\right\}.$$

As $n \to \infty$,

$$E(M) \sim \theta \log n, \ \mathrm{var}(M) \sim \theta \log n.$$

The distribution of M can be expressed as a Poisson mixture

$$P_n(z) = \prod_{j=1}^{n-1} \left(1 - \frac{(z-1)\theta}{j}\right)^{-1}$$

$$= \prod_{j=1}^{n-1} \frac{j}{j + (1-z)\theta}$$

$$= \frac{(n-1)! \, \Gamma\big(1 + (1-z)\theta\big)}{\Gamma\big(n + (1-z)\theta\big)}$$

$$= (n-1)B\big(n-1, 1 + (1-z)\theta\big)$$

$$= (n-1) \int_0^1 x^{n-2}(1-x)^{(1-z)\theta} dx$$

$$P_n(z) = (n-1) \int_0^1 x^{n-2} G(z; x) dx,$$

where

$$G(z; x) = \exp\left\{(z-1)\big(-\theta \log(1-x)\big)\right\}.$$

Thus

$$P(M = m) = (n-1) \int_0^1 x^{n-2} \frac{\lambda(x)^m}{m!} e^{-\lambda(x)} dx,$$

where $\lambda(x) = -\theta \log(1-x)$.

The TMRCA of a sample of Y-chromosomes.

Robert Dorit, Hiroshi Akashi,
and Walter Gilbert.
Absence of polymorphism at the ZFY locus on
the human Y-chromosome.
Science, **268**, 1183–1185, May 1995.
38 individuals were observed to have no variation
(*ie* no mutations) at the ZFY locus.

Problem. Estimate the TMRCA of the sample of
chromosomes, given no variation in the sample.
Their estimate is 270,000 years, with a 95% con-
fidence interval of (0,800,000) years.
Their incorrect equation used for estimating the
TMRCA:

$$P\big(\text{No mutation} \mid \text{TMRCA} = T\big)$$

$$= \prod_{i=2}^{38} \frac{i-1}{i-1+\mu T},$$

where μ is the mutation rate.

Coalescent theory approach.

The probability of no mutation given $T_j = t$, the time while j ancestors is

$$e^{-j \cdot \frac{\theta}{2} t}.$$

The density of T_j is

$$\binom{j}{2} e^{-\binom{j}{2}t}, \ t > 0,$$

so the density of T_j given no mutations is proportional to

$$e^{-j \cdot \frac{\theta}{2} t} \cdot e^{-\binom{j}{2}t}$$

$$= e^{-\frac{j(j+\theta-1)}{2}t}, \ t > 0.$$

ie. The conditional distribution is exponential with mean $2/j(j + \theta - 1)$.

Assuming a 20 year generation and $\mu = 1.96 \times 10^{-5}$,

$$E(\text{TMRCA} \mid \text{no mutation})$$

$$= 20N \sum_{j=2}^{38} \frac{2}{j(j + 2N\mu - 1)}$$

where N is the male effective population size.

N	E(TMRCA|no mutation)
2,500	93,000
5,000	177,000
10,000	324,000
20,000	563,000

Dorit *et al*'s analysis drew comments from several theoretical population genetics groups in *Science.*

For a more extensive analysis see:

P. Donnelly, S. Tavaré, D. Balding and R.C. Griffiths.

Estimating the age of the common ancestor of men from the ZFY intron.

Science (1996), **272**, 1357–1359.

There is a simple algorithm to simulate the TMRCA of a sample conditional on k mutations in a sample.

The mean can then be found from the simulated data.

This algorithm is in the paper:

S. Tavaré, D. Balding, R. C.Griffiths and P. Donnelly.

Inferring coalescence times from DNA sequence data.

Genetics (1997), **145**, 505–518.

1. Simulate $\{T_j, j = n, \ldots, 2\}$, independent exponential random variables with parameters $\binom{j}{2}$, $j = n, \ldots, 2$.

2. Evaluate the TMRCA,

$$W = \sum_{j=2}^{n} T_j$$

and the total edge length in the tree

$$L_n = \sum_{j=2}^{n} j T_j.$$

3. Keep W generated with probability u, defined by

$$u = \frac{\text{Poisson}\,(k, L_n\theta/2)}{\text{Poisson}\,(k, k)}$$

otherwise discard W generated and go to step 1.

It is possible to replace step 3 with a MCMC step.

3′. Keep W generated with probability

$$\min(1, u/u')$$

where
$$u = \text{Poisson } (k, L_n\theta/2)$$

and u' is a similar, from the previous run.
If W is not accepted, keep W' from the previous run instead.

In the MCMC approach TMRCA generated values are possibly repeated and a typical realization is like

$$W_1, W_1, \ldots, W_2, W_2, \ldots$$

A burnin period is selected where the simulation is run for some time before then sampling with gaps of a fixed size from the sequence.
In this problem it is reasonable to have a burnin of (say) 1000 and gap of 100 between sampling.

Variable population sized models.

The population size at the current time is N_0, with the size at time t back

$$N(t) = N_0 \nu(t).$$

Coalescence times $\{T_j\}$.
Let

$$S_j = T_n + \cdots T_j, \ S_{n+1} = 0,$$

$\{S_j\}$ forms a reverse Markov process,

$$P(S_j > s_j \mid S_{j+1} = s_{j+1})$$
$$= \exp\left(-\binom{j}{2}\int_{s_{j+1}}^{s_j}\frac{dt}{\nu(t)}\right)$$

A coalescent process with growth can be coupled in distribution with a constant sized process. Using the notation $\{S_j^\nu\}$ for variable population size coalescent times and $\{S_j^\circ\}$ for constant size, the coupling is that

$$S_j^\circ = \int_0^{S_j^\nu}\frac{dt}{\nu(t)}.$$

Simulation.

For $j = n, n-1, \ldots, 2$ solve for S_j (omitting the superscript ν) in

$$\int_{s_{j+1}}^{s_j} \frac{dt}{\nu(t)} = -\binom{j}{2}^{-1} \log U_j,$$
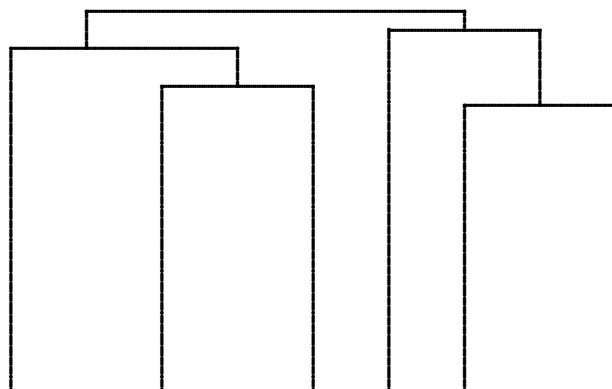
where $\{U_j\}$ are *i.i.d.* uniform random variables on $(0, 1)$.

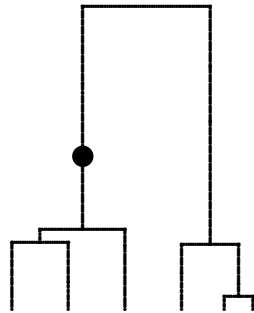*Exponential growth:* $\nu(t) = e^{-\beta t}$

$$\frac{e^{\beta s_j} - e^{\beta s_{j+1}}}{\beta} = -\binom{j}{2}^{-1} \log U_j,$$

$$s_j = \beta^{-1} \log \left(e^{\beta s_{j+1}} - \beta \binom{j}{2}^{-1} \log U_j \right).$$

Exponential growth shortens coalescence times and makes coalescent trees `star like`.

Mutant genes in a sample.

A general binary coalescent tree has continuous coalescence times $T_n, \ldots, T_2$ with an exchangeable coalescence structure such that any pair of edges while k edges have an equal probability $\binom{k}{2}^{-1}$ of coalescing.
Denote

$$p_{nk}(b) = \frac{\binom{n-b-1}{k-2}}{\binom{n-1}{k-1}}$$

as the probability that an edge while k ancestor lines subtends b descendants in the sample.
If a mutation falls on such an edge while k ancestors it will be represented in b genes in the sample.

An urn representation.

$\times k$
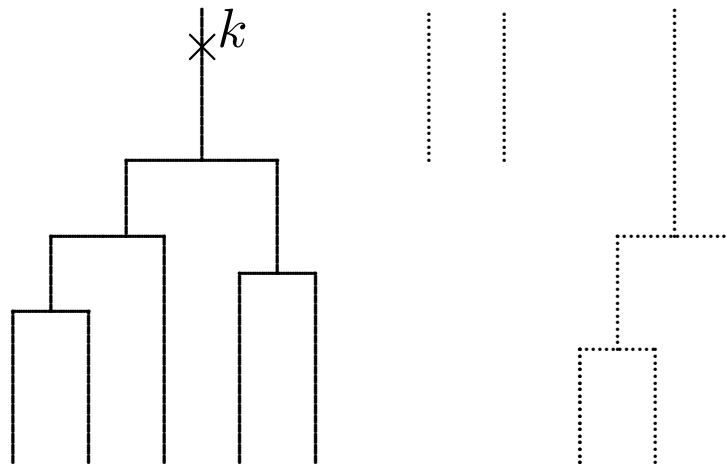
Classical de Finetti urn
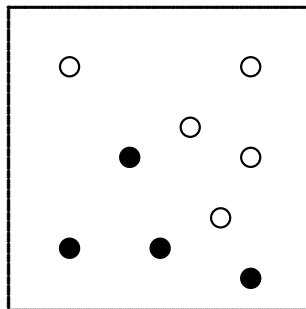
(a) Put one black ball and k-1 white balls in an urn.

(b) At each trial draw a ball at random and re-place with an additional ball of the colour chosen.

(c) Stop when n balls.

Forward in time branching in the subtree
is equivalent to drawing a black ball.

The distribution of the number of black balls is
the same as the number of descendents in the
subtree under the mutation.

Problem. A mutation is observed in b out of n genes in a sample, where $0 < b < n$.

What is the probability distribution of the number of copies of b?

It is helpful in this problem to label mutations as independent uniform random variables on [0,1]. Then the probability that a mutation has a label in $(x, x + h)$ is h.

Answer. Let $C_h = C(x, b, h)$ denote the event that there is a mutation with label U in the interval $(x, x + h) \subset (0, 1)$ that subtends b copies in the sample, and let I_k denote the event that this mutation arises while k ancestors.

Then with $\mathbf{T} = (T_n, \ldots, T_2)$,

$$P(C_h \mid \mathbf{T})$$

$$= \sum_k P(C_h, I_k \mid \mathbf{T})$$

$$= \sum_k p_{nk}(b) P(I_k, U \in (x, x+h) \mid \mathbf{T})$$

$$= \sum_k p_{nk}(b) \left(k T_k \frac{\theta}{2} h + o(h) \right)$$

Averaging over the distribution of $\mathbf{T}$,

$$P(C_h) \sim \frac{\theta h}{2} \cdot \sum_k k p_{nk}(b) E(T_k), \text{ as } h \to 0.$$

Summing over b, the probability that there is a mutation with label in $(x, x+h)$ is

$$\sim \frac{\theta h}{2} \cdot \sum_k k E(T_k), \text{ as } h \to 0.$$

The probability that a particular mutation has b copies is thus

$$q_{nb} = \frac{\sum_{k=2}^n k p_{nk}(b) E(T_k)}{\sum_{k=2}^n k E(T_k).}, \ 0 < b < n.$$

Exercise. Show that the expected number of mutant copies of a mutation in the distribution

$\{q_{nb}\}$ is

$$\mu = \frac{n \sum_{k=2}^{n} E(T_k)}{\sum_{k=2}^{n} k E(T_k)}$$

Constant population size.

$$E(T_k) = \binom{k}{2}^{-1}, \ k = n, \ldots, 2.$$

As an exercise

$$\sum_{k=2}^{n} k p_{nk}(b) E(T_k) = 2b^{-1},$$

and

$$q_{nb} = \frac{b^{-1}}{\sum_{j=1}^{n-1} j^{-1}}, \ , 1 \le b \le n - 1.$$

The mean number of mutations in the distribution is

$$\mu = \frac{n - 1}{\sum_{j=1}^{n-1} j^{-1}}.$$

As $n \to \infty$,

$$\mu \sim \frac{n}{\log n}, \ \sigma^2 \sim \frac{n^2}{2 \log n}.$$

The age of a mutation.

Problem. A mutation is observed to occur in b genes out of n in a sample.

What is the expected age of the mutation?

Answer. The solution is similar to the derivation of q_{nb}.

Denote the age of the mutation by ξ_{nb}. Given a mutation occurs while k ancestors its age is

$$UT_k + T_{k+1} + \cdots + T_n,$$

where U is a uniform random variable independent of $\{T_j\}$. The mean age is therefore

$$E(\xi_{nb}) = \frac{\sum_{k=2}^{n} k p_{nk}(b) E\left(T_k\left(\frac{1}{2}T_k + \cdots + T_n\right)\right)}{\sum_{k=2}^{n} k p_{nk}(b) E(T_k)}.$$

In a constant size population the denominator in the formula for $E(\xi_{nb})$ is $2b^{-1}$, and

$$E\Big(T_k\big(\tfrac{1}{2}T_k + T_{k-1} + \cdots + T_n\big)\Big)$$

$$= \frac{1}{2} \cdot 2 \cdot \binom{k}{2}^{-2} + \binom{k}{2}^{-1} \sum_{j=k+1}^{n} \binom{j}{2}^{-1}$$

$$= \binom{k}{2}^{-1} \cdot 2\left[\frac{1}{k-1} - \frac{1}{n}\right]$$

$$= 2\binom{k}{2}^{-1} \frac{n-k+1}{n(k-1)}.$$

Now

$$kp_{nk}(b) = k \cdot \frac{(n-b-1)!}{(k-2)!(n-b-k+1)!}$$

$$\cdot \frac{(k-1)!(n-k)!}{(n-1)!}$$

$$= k(k-1)\binom{n-k}{b-1} \cdot b^{-1} \binom{n-1}{b}^{-1},$$

so

$$E(\xi_{nb}) = 2\binom{n-1}{b}^{-1} \sum_{j=2}^{n} \binom{n-j}{b-1} \frac{n-j+1}{n(j-1)}$$

$$= 2b(n-b)^{-1} \sum_{j=b+1}^{n} j^{-1}.$$

The last identity is not immediate.
Exercise: Prove that it is true.

The age of an allele of frequency x in the popu-lation.

To obtain a population analogue of $E(\xi_{nb})$ take
the limit as $n \to \infty$, $b \to \infty$, while $b/n \to x$.

$$E(\xi_{nb}) = 2b(n-b)^{-1} \sum_{j=b+1}^{n} j^{-1}$$

$$= 2(b/n)(1-b/n)^{-1} \sum_{j=b+1}^{n} (j/n)^{-1} \cdot n^{-1}$$

$$\to 2x(1-x)^{-1} \int_{x}^{1} u^{-1} du$$

$$= -2x(1-x)^{-1} \log(x).$$

This is a well known classical result, proved in a different way before the coalescent was developed in:

Kimura, M and Ohta, T (1973). The age of a neutral mutant persisting in a finite population. *Genetics* **75**, 199–212.

The coalescent approach is in:

Griffiths, R. C. and Tavaré (1998). The age of a mutation in a general coalescent tree. *Stochastic models* **14**, 273–295,

Wiuf, C and Donnelly, P. (1999) Conditional genealogies and the age of a neutral mutant. *Theor. Popul. Biol.* **56**, 183–201,

and a diffusion theory approach is in:

Griffiths, R. C. (2003). The frequency spectrum of a mutation, and its age, in a general diffusion model. *Theor. Popul. Biol.* 64, 241-251.

The infinitely-many-alleles model.

Each mutation that occurs in the coalescent tree is assumed to be new.
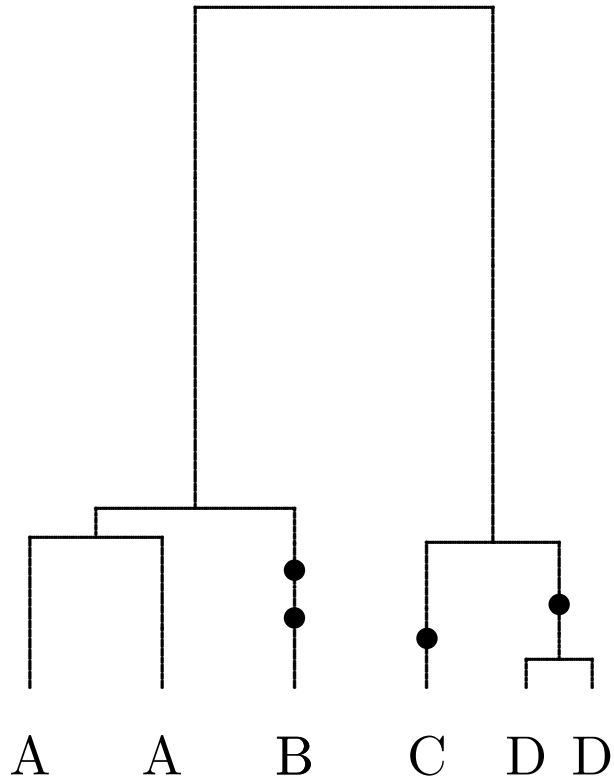
With an arbitary labelling the types in the sample are A A B C D D.

The probability distribution of the sample configuration is known as the *Ewens' sampling formula.*

Denote $\alpha(j), 1 \leq j \leq n$ as the number of alleles with j representatives. The total sample size is n, so

$$\sum_{j=1}^{n} j\alpha(j) = n.$$

The number of alleles (different types of gene) is

$$K = \sum_{j=1}^{n} \alpha(j).$$

In the example tree $\alpha(1) = 2, \alpha(2) = 2$, and $K = 4$.

The probability distribution for the configuration is

$$P(\{\alpha(j)\}) =$$

$$\frac{n!\theta^k}{\alpha(1)! \cdots \alpha(n)! \cdot 1^{\alpha(1)} \cdots n^{\alpha(n)} \cdot \theta_{(n)}},$$

with $\sum_{j=1}^{n} j\alpha(j) = n$, $k = \sum_{j=1}^{n} \alpha(j)$, and

$$\theta_{(n)} = \theta(\theta+1) \cdots (\theta+n-1).$$

K, the number of alleles in the sample, is a sufficient statistic for θ.

Non-mutant lines of descent.

Let $A_n^\theta(t)$ be the number of non-mutant lines at time t back. Lines can be `lost` by coalescence at rate $k(k-1)/2$ while k ancestors, or by mutation at rate $\theta k/2$.

Thus $\{A_n^\theta(t), t \geq 0\}$ is a death process with death rates

$$\mu_k = \frac{k(k-1+\theta)}{2} \quad k = n, \ldots, 1.$$

The number of alleles.

Let $X_1, \ldots, X_n$ be indicator variables as to whether the ith line is lost by mutation or coalescence.

$$X_i = \begin{cases} 1 & \text{if lost by mutation,} \\ 0 & \text{if lost by coalescence.} \end{cases}$$

Then

$$K = 1 + X_2 + \cdots + X_n.$$

The last line must be lost by mutation so $X_1 = 1$. $\{X_j\}$ are independent random variables and

$$P(X_j = 1) = \frac{\theta j/2}{j(j-1)/2 + \theta j/2}$$

$$= \frac{\theta}{j + \theta - 1}$$

The mean and variance of X_j are

$$E(X_j) = \frac{\theta}{j + \theta - 1}$$

$$\mathrm{var}(X_j) = \frac{\theta(j-1)}{(j + \theta - 1)^2},$$

so the mean and variance of K are

$$E(K) = 1 + \theta \sum_{j=2}^{n} \frac{1}{j + \theta - 1}$$

$$\mathrm{var}(K) = \theta \sum_{j=2}^{n} \frac{j-1}{(j + \theta - 1)^2}.$$

As $n \to \infty$, $E(K) \sim \theta \log n$, $\mathrm{var}(K) \sim \theta \log n$. The *pgf* of K is

$$\prod_{j=1}^{n} \left(\frac{j-1}{j + \theta - 1} + \frac{z\theta}{j + \theta - 1} \right)$$

$$= \frac{(\theta z)_{(n)}}{\theta_{(n)}}$$

To invert the *pgf* of K,

$$\frac{(\theta z)_{(n)}}{\theta_{(n)}},$$

consider the expansion

$$\phi(\phi - 1)\cdots(\phi - n + 1) = \sum_{k=1}^{n} S_k^n \phi^k,$$

where ϕ is an arbitary variable and $\{S_k^n\}$ are Stirling numbers of the 1st kind.

$$P(K = k) = \frac{\theta^k}{\theta_{(n)}} \cdot |S_k^n|, k = 1, 2, \ldots, n.$$

The maximum likelihood estimate of θ in a sample of n is found from solving

$$\frac{\partial}{\partial\theta}\left\{k\log\theta - \sum_{j=1}^{n}\log(\theta + j - 1)\right\} = 0,$$

$$\frac{k}{\hat{\theta}} = \sum_{j=1}^{n}\frac{1}{\hat{\theta} + j - 1},$$

$$k = \sum_{j=1}^{n}\frac{\hat{\theta}}{\hat{\theta} + j - 1},$$

That is, the MLE is the 1st moment estimate of θ.

An urn model representation for the allele sample.
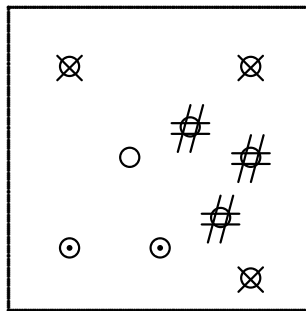
1. Start with 1 white ball of mass θ in the urn.

2. Select a ball from the urn. If it is white return it with a ball of a new colour, if not add a ball of mass 1 of the same colour as the ball drawn.

3. Stop when n non-white balls and randomly label them $1, 2, \ldots, k$ if k different colours.

The urn model is known as Hoppe's urn model.

The identification of the coalescent model with the urn representation is made by considering what happens at the 1st event back in time in the coalescent tree.
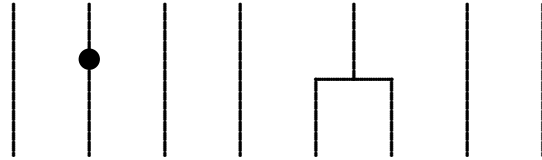

The probability that the 1st event back in time was a coalescence is $(n-1)/(n-1+\theta)$, and a mutation $\theta/(n-1+\theta)$.

These are the probabilities of choosing an existing coloured ball, or the ball of mass θ from the urn when it has $n-1$ coloured balls plus the ball of mass θ.

Choosing a ball of an existing colour in the urn and adding another is identified with branching in the coalescent tree.

Ewens' sampling formula can be proved by induction from the urn model.

Coalescent tree and Gene tree.

Coalescent tree.
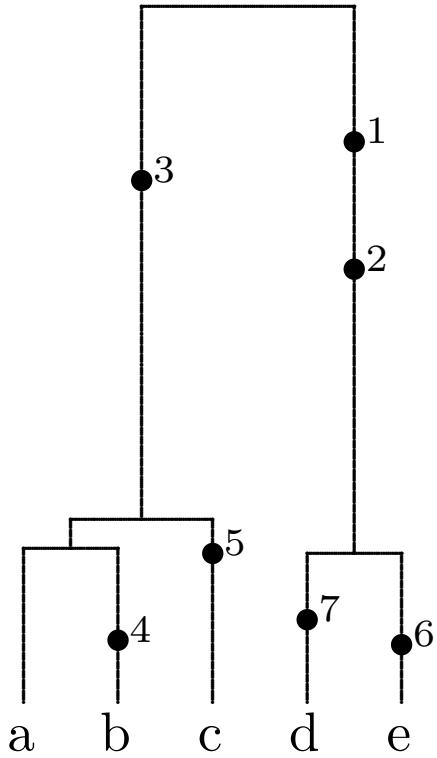
Gene tree.

Mutation pattern on sequences

	1	2	3	4	5	6	7
a			✕				
b			✕	✕			
c			✕		✕		
d	✕	✕					✕
e	✕	✕				✕	
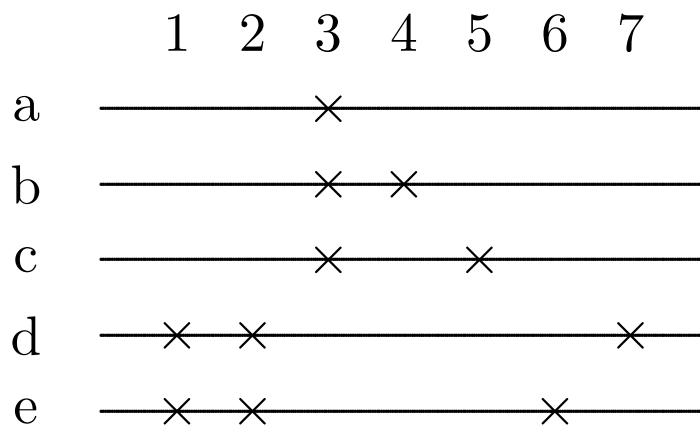
Infinitely-many-sites model.

Mutations occur at positions on the DNA sequences never before mutant.

Every mutation occurring in the coalescent tree on an edge occurs in all genes subtended below the edge.

A site with mutant and ancestor types is called a `segregating site`.

The number of segregating sites in a sample of DNA sequences
= the number of mutations on the coalescent tree.
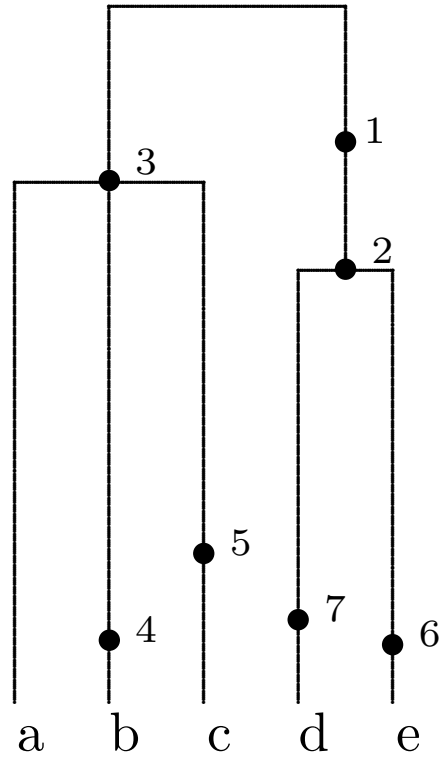
In a sample of n sequences with s segregating sites the $n \times s$ `incidence matrix` S of mutations on sequences is obtained by letting

$$s_{ij} = \begin{cases} 1 & \text{if sequence } i \text{ has mutation } j \\ 0 & \text{otherwise.} \end{cases}$$

In the example coalescent tree the incidence matrix is

	1	2	3	4	5	6	7
a	0	0	1	0	0	0	0
b	0	0	1	1	0	0	0
c	0	0	1	0	1	0	0
d	1	1	0	0	0	0	1
e	1	1	0	0	0	1	0

Gene tree.
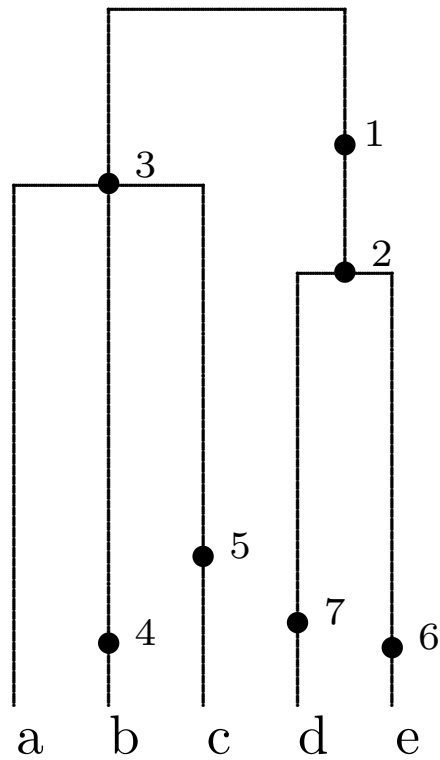
A `gene tree` has mutations as vertices and describes the mutation history of the sequences. Paths to the root (denoted as 0) in the example gene tree:

a	3	0		
b	4	3	0	
c	5	3	0	
d	7	2	1	0
e	6	2	1	0

DNA sequences and Gene trees.

In a sample of n sequences suppose there are s segregating sites, corresponding to s mutations. Label the mutations $1, 2, \ldots, s$ and let $O_1, \ldots, O_s$ be the sets of sequences containing mutations $1, 2, \ldots, s$.

Example.

	1	2	3	4	5	6	7
1	0	0	1	0	0	0	0
2	0	0	1	1	0	0	0
3	0	0	1	0	1	0	0
4	1	1	0	0	0	0	1
5	1	1	0	0	0	1	0

$$O_1 = \{4, 5\}, O_2 = \{4, 5\}, O_3 = \{1, 2, 3\},$$
$$O_4 = \{2\}, O_5 = \{3\}, O_6 = \{5\}, O_7 = \{4\}.$$

The sets $O_1, \ldots, O_s$ are partially ordered by inclusion, that is, for $i \neq j$ either

$$O_i \subset O_j, O_j \subset O_i, \text{ or } O_i \cap O_j = \phi.$$

This is easy to see from a tree.
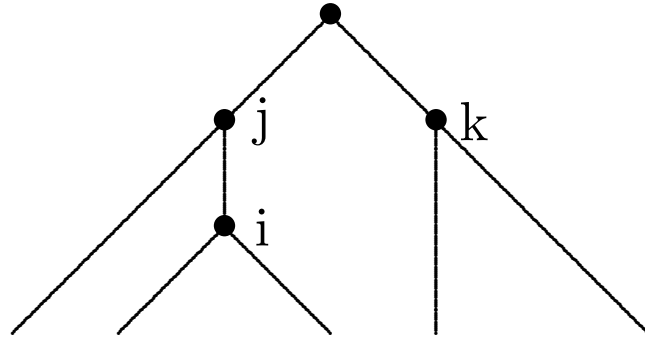
$O_i \subset O_j, \; O_j \cap O_k$ is empty.

Gusfield's algorithm

1. Represent duplicate columns in the incidence matrix as a single column with a label corresponding to the identical columns, for example (1,6,8).

2. Considering each column as a binary number, sort the numbers into decreasing order, with the largest number in column 1.

3. Construct paths from the leaves to the root in the gene tree by labelling nodes by mutation column labels, and reading vertices in paths from the right to the left where 1's occur in rows.

   Gusfield, D.(1991). Efficient algorithms for inferring evolutionary trees. *Networks,* **21**, 19–28.

DNA sequences
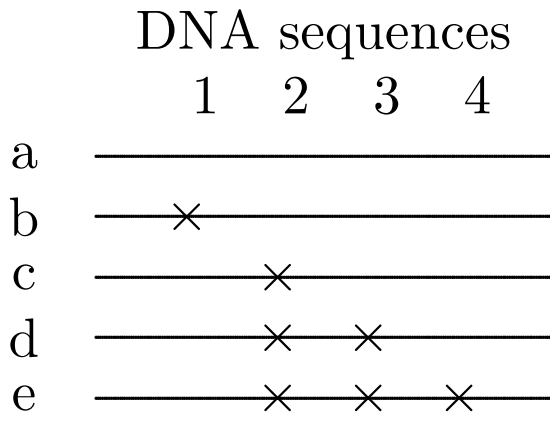
	1	2	3	4
a				
b	✕			
c		✕		
d		✕	✕	
e		✕	✕	✕

Incidence matrix

	1	2	3	4
a	0	0	0	0
b	1	0	0	0
c	0	1	0	0
d	0	1	1	0
e	0	1	1	1

Paths to the root

a	0			
b	1	0		
c	2	0		
d	3	2	0	
e	4	3	2	0

Hammer's Y tree.

56
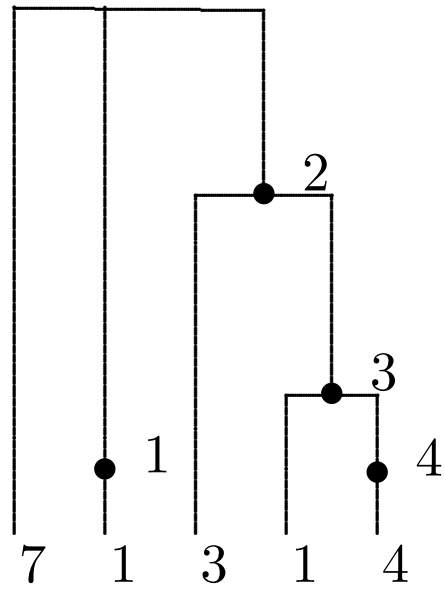
A recent common ancestry
for human Y chromosomes
Michael F. Hammer, *Nature* 1995.

Three mutations ● and an insertion ⊗.
D and E do not contain the insertion.

Take the mutation rate to be $\theta = 0.97$.
Calculated from 2.6×10^3 bases, at a rate of 1.9×10^{-9} per base per year, 20 year generation, $N_e = 4900$.
Sample size of $n = 16$ sequences.

Mike Hammer's inferred tree:

Theorem. The configuration of mutations on sequences is equivalent to a gene tree.

Proof. It is clear how to obtain an incidence matrix of mutations on sequences from the gene tree.

To construct a gene tree from the mutation configuration on the sequences Gusfield's algorithm is used.

An induction proof on the number of mutations shows that the algorithm constructs a gene tree. Without loss of generality assume that there are no duplicate columns.

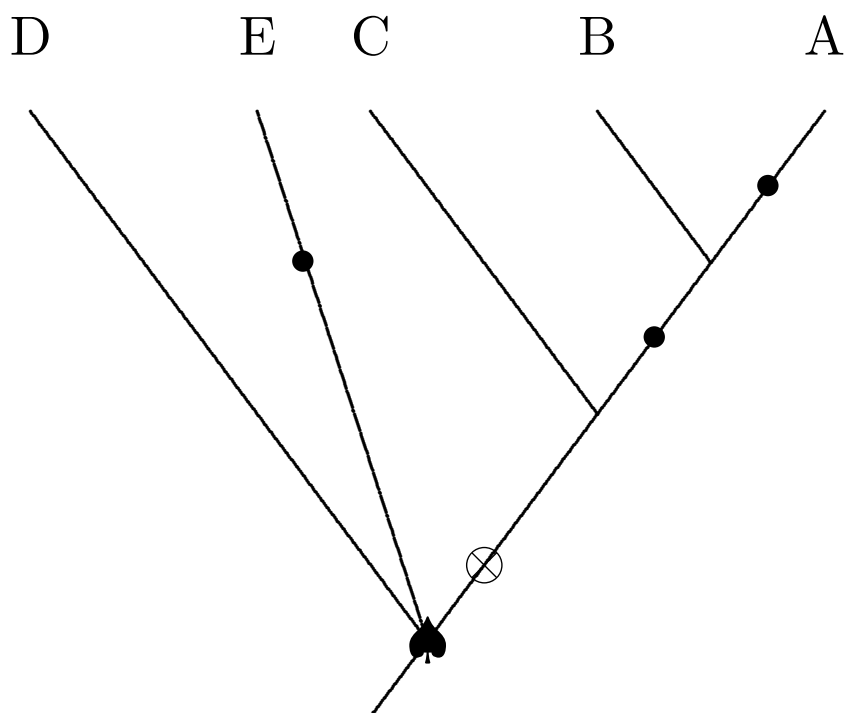If there is one mutation labelled 1 only in (say) the 1st k sequences the algorithm gives paths to the root (labelled 0) of:

$$\begin{cases} 1\ 0 & \text{for the 1st } k \text{ sequences} \\ 0 & \text{for the last } n - k \text{ sequences.} \end{cases}$$

The incidence matrix is

$$\begin{matrix} 1 \\ \vdots \\ 1 \\ 0 \\ \vdots \\ 0 \end{matrix}$$

Suppose there are s mutations.
In the sorted order as binary numbers if

$$\text{column } i < \text{ column } j$$

then $O_i \subset O_j$ or $O_i \cap O_j = \phi$.

Assume a unique tree is constructed from the 1st $s - 1$ columns.
If $O_s \cap O_j = \phi, j < s$, then for each row containing a 1 in column s add s to the vertex path to the root.

	Tree with $s-1$ mutations		Tree with s mutations
	$\vdots$		$\vdots$
	0	s	0
	0	s	0
	$\vdots$		$\vdots$
	0		0

Columns j and s in the incidence matrix have the form

j	s
1	0
$\vdots$	$\vdots$
1	0
0	1
$\vdots$	$\vdots$
0	1
$\vdots$	$\vdots$
0	0

If instead $O_s \cap O_j = \phi$, $j = s-1, \ldots, k+1$ and $O_s \subset O_k$ add s to the vertex path.

Tree with $s-1$ mutations	Tree with s mutations
$\vdots$	$\vdots$
$k \ldots 0$	$s\ k \ldots 0$
$k \ldots 0$	$s\ k \ldots 0$
$\vdots$	$\vdots$
$k \ldots 0$	$k \ldots 0$

Three rows k, j, s in the incidence matrix have the form

k	j	s
1	1	0
1	1	0
0	1	0
1	0	1
1	0	1
1	0	0

Given the $(s-1)$-tree the s-tree is unique. This completes the induction proof.

Duplicate columns in the incidence matrix.

	1	2	3	4	5	6	7
1	0	0	1	0	0	0	0
2	0	0	1	1	0	0	0
3	0	0	1	0	1	0	0
4	1	1	0	0	0	0	1
5	1	1	0	0	0	1	0

Columns 1 and 2 are identical.
The gene tree is unique up to a permutation of labels 1 and 2.

Mutation pattern on sequences
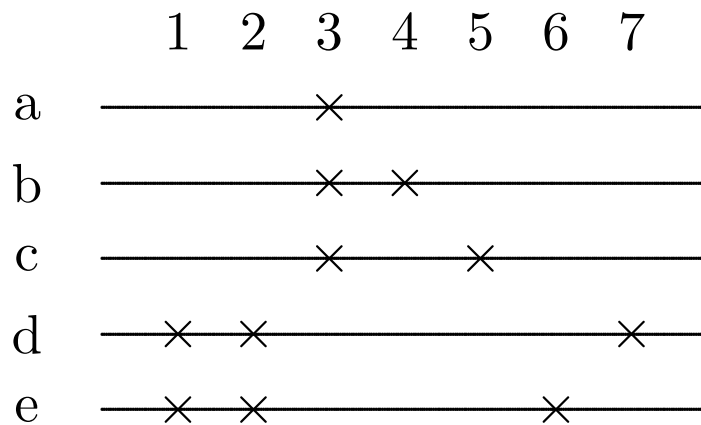
Unrooted trees.

If it is not known which type is the mutant base and which is the ancestor base then the tree is unrooted. Vertices are sequences in the unrooted tree. There can be inferred sequences in the tree. If there are s mutations then there are $s + 1$ possible rooted trees, to the left or right of each mutation.

Coalescent tree.

Unrooted tree

Mitochondrial DNA sample

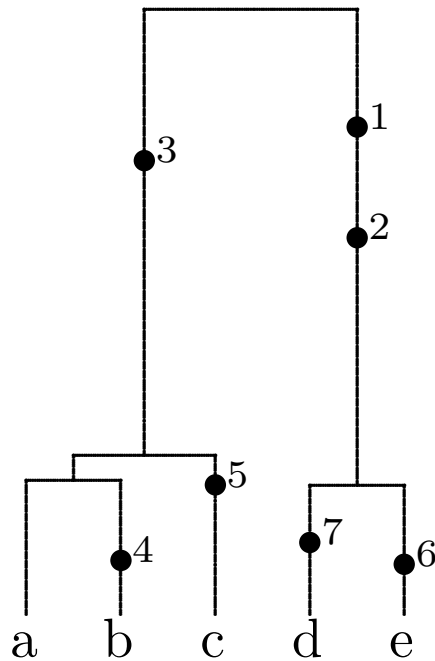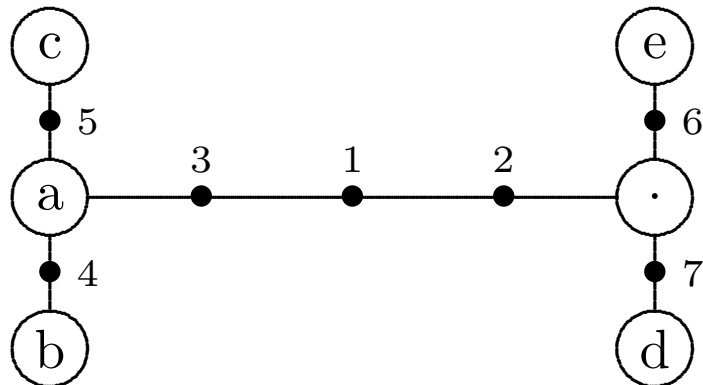|	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	Freq
a	A	G	G	A	A	T	C	C	T	C	T	T	C	T	C	T	T	C	2
b	A	G	G	A	A	T	C	C	T	T	T	T	C	T	C	T	T	C	2
c	G	A	G	G	A	C	C	C	T	C	T	T	C	C	C	T	T	T	1
d	G	G	A	G	A	C	C	C	C	C	T	T	C	C	C	T	T	C	3
e	G	G	G	A	A	T	C	C	T	C	T	T	C	T	C	T	T	C	19
f	G	G	G	A	G	T	C	C	T	C	T	T	C	T	C	T	T	C	1
g	G	G	G	G	A	C	C	C	T	C	C	C	C	C	C	T	T	T	1
h	G	G	G	G	A	C	C	C	T	C	C	C	T	C	C	T	T	T	1
i	G	G	G	G	A	C	C	C	T	C	T	T	C	C	C	C	C	T	4
j	G	G	G	G	A	C	C	C	T	C	T	T	C	C	C	C	T	T	8
k	G	G	G	G	A	C	C	C	T	C	T	T	C	C	C	T	T	C	5
l	G	G	G	G	A	C	C	C	T	C	T	T	C	C	C	T	T	T	4
m	G	G	G	G	A	C	C	T	T	C	T	T	C	C	C	T	T	C	3
n	G	G	G	G	A	C	T	C	T	C	T	T	C	C	T	T	T	C	1

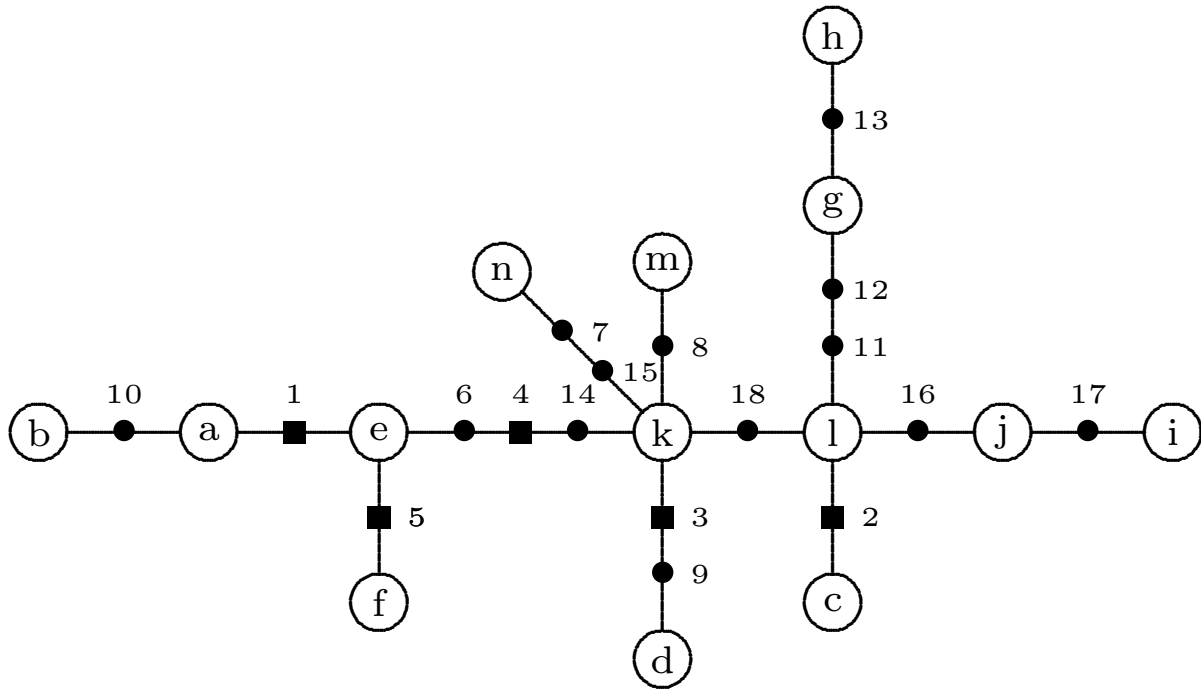Ward, R. H. Frazier, B. L., Dew, K. and Paabo, S. (1991)
Extensive mitochondrial diversity within a single Amerindian tribe.
Proc. Nat. Acad. Sci. USA **88** 8720-8724.
North American Indian tribe,
the Nuu-Chah-Nulth from Vancouver Island.
$N = 600$ (women).

Unrooted Nuu-Chah-Nulth tree

● pyrimidine sites; ■ purine sites

Compatibility of mutations with the point mutations assumption

An $n \times s$ 0-1 matrix is compatible with a gene tree if and only if no pattern

$$
\begin{array}{cc}
0 & 1 \\
1 & 0 \\
1 & 1
\end{array}
$$

occurs in any two columns and three rows.

Necessity Label the three sequences 1,2,3, the mutations 1,2 and consider O_1, O_2 the sets of sequences containing mutations 1 and 2. Assume the three sequences and two mutations form a gene tree.

We know that O_1, O_2 must be ordered by inclusion. However $O_1 = \{2, 3\}$ and $O_2 = \{1, 3\}$, and the sets are not ordered by inclusion because $O_1 \cap O_2 = \{3\}$.

This is a contradiction, proving the necessity.

Sufficiency. Suppose there is no such pattern

$$
\begin{array}{cc}
0 & 1 \\
1 & 0 \\
1 & 1
\end{array}
$$

An induction proof is given that it is possible to construct a gene tree on the number of mutations.

It is trivially true for 1 mutation.

Suppose the condition is sufficient for a tree with up to $s - 1$ mutations. Let O_j be the set of sequences containing mutation j, $j = 1, \ldots, s$.

If $\{O_j\}$ is partially ordered by inclusion it is possible to construct a gene tree from a previous theorem.

If O_s doesn't fit into the partial ordering, then for some $i < s$, $O_i \cap O_s \neq \phi, O_i, O_s$ and there exist distinct sequences labelled a, b, c such that

$$a \in O_i, a \notin O_s$$
$$b \in O_s, b \notin O_i$$
$$c \in O_i \cap O_s.$$

The mutation configuration at sites i and s is

	i	s
a	1	0
b	0	1
c	1	1

This is a contradiction to there being no such pattern, therefore O_s must fit into the partial ordering.

The induction proof is then complete, and so is the full proof.

If the mutant and ancestor base types are not known at sites, then an $n \times s$ 0-1 matrix is compatible with a point mutation model if and only if no two columns and four rows have a mutation pattern of

$$
\begin{matrix}
0 & 0 \\
0 & 1 \\
1 & 0 \\
1 & 1
\end{matrix}
$$

Furthur, if the columns are coded so that 0 is the most frequent type, then an unrooted tree exists if and only if this coded matrix represents a rooted tree with the most frequent base at each site.

(i) Suppose that the pattern above does occur. The pattern is invariant under toggling 0 and 1's in columns, so whichever pattern under toggling potentially represents a rooted tree it must contain

$$
\begin{matrix}
0 & 1 \\
1 & 0 \\
1 & 1
\end{matrix}
$$

Therefore the matrix configuration is not compatible with any rooted tree.

(ii) Suppose the pattern does not occur and consider constructing a rooted tree from a mutation configuration where 0 and 1's in columns are toggled so that 0 is the most or equally frequent. Suppose that this configuration does not represent a rooted tree. Then there must be a pattern

$$
\begin{array}{cc}
0 & 1 \\
1 & 0 \\
1 & 1
\end{array}
$$

occurring in the matrix.

However the matrix was chosen so that 0 was the most or equally frequent in each column.

There are two 1's in each column in the pattern, so adding rows in the matrix with $0, 1$; $1, 1$ or $1, 0$ configurations increases the number of 1's in a column.

Therefore there must be a row with pattern $0, 0$ in these columns and a pattern

$$
\begin{array}{cc}
0 & 1 \\
1 & 0 \\
1 & 1 \\
0 & 0
\end{array}
$$

This is a contradiction, so the rooted tree corresponding to 0 (the wild type mutation) being the most frequent exists, and thus does the unrooted tree.
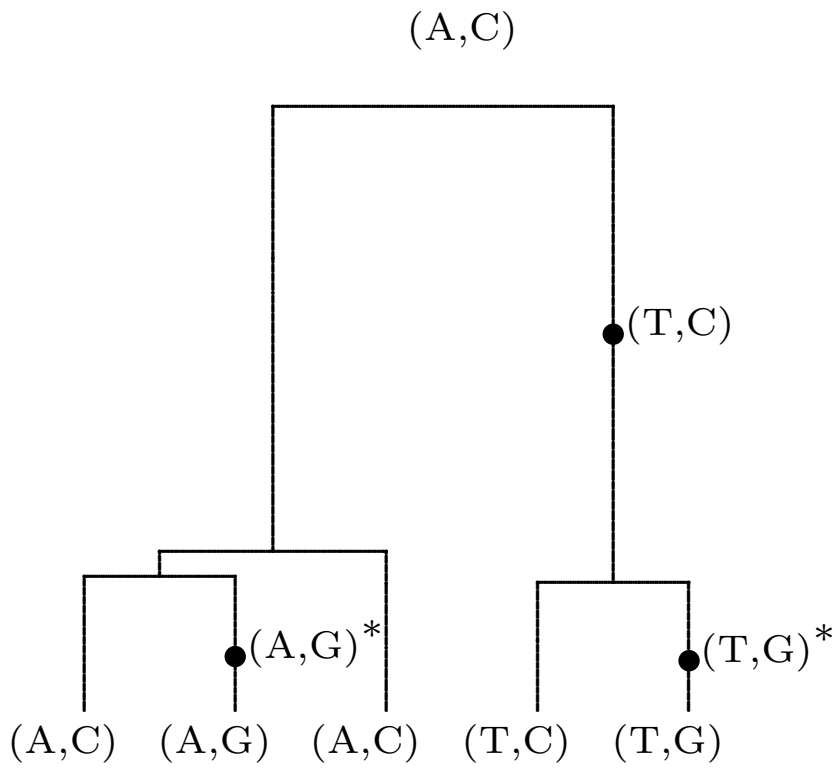
Parallel mutation C → G
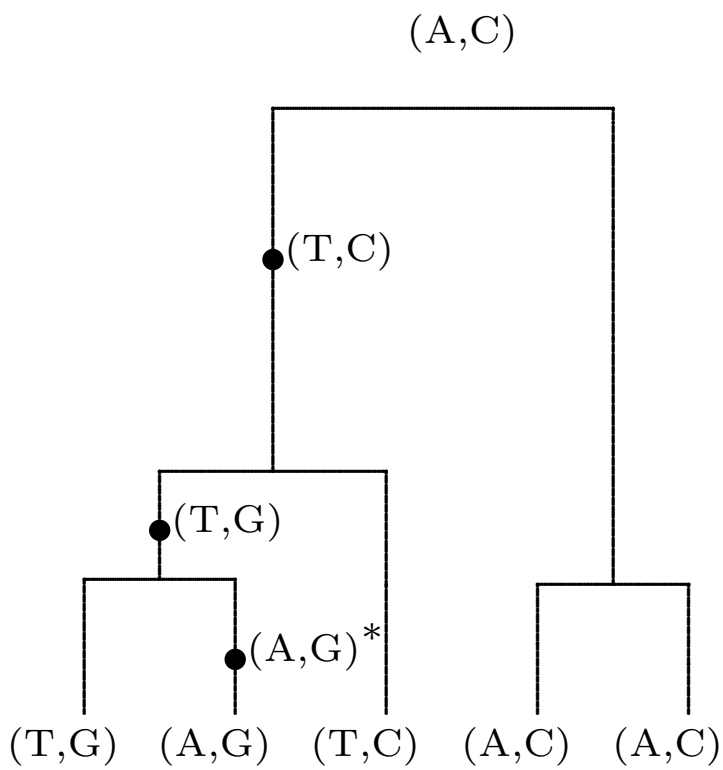
Back mutation T → A

Probability distribution of gene trees $(T, \mathbf{n})$

```
25 :  1   2   0
 1 :  3   1   2  0
16 :  4   0
 3 :  5   6   7  8  0
 7 :  9  10   5  6  7  8  0
 1 : 11  10   5  6  7  8  0
 4 : 12  13  10  5  6  7  8  0
```

Gene tree T contains sequences of mutation
paths from the leaves to the root.
$\mathbf{n}$ contains sequence multiplicities.
$p(T, \mathbf{n})$ is the probability of a gene tree.

Melanesian β-globin sequences.

Site	1	2	3	4	5	6	7	8	9	10	11	12	13	
Root	T	T	T	A	T	C	T	C	T	C	G	G	C	
Lineage														Freq
a	G	G	T	A	T	C	T	C	T	C	G	G	C	25
b	G	G	C	A	T	C	T	C	T	C	G	G	C	1
c	T	T	T	T	T	C	T	C	T	C	G	G	C	16
d	T	T	T	A	C	T	C	A	T	C	G	G	C	3
e	T	T	T	A	C	T	C	A	C	G	G	G	C	7
f	T	T	T	A	C	T	C	A	T	G	A	G	C	1
g	T	T	T	A	C	T	C	A	T	G	G	C	T	4

Melanesian β-globin tree

73

Melanesian β-globin sequences.

Site	1	2	3	4	5	6	7	8	9	10	11	12	13	
Root	T	T	T	A	T	C	T	C	T	C	G	G	C	
Lineage														Freq
a	G	G	T	A	T	C	T	C	T	C	G	G	C	25
b	G	G	C	A	T	C	T	C	T	C	G	G	C	1
c	T	T	T	T	T	C	T	C	T	C	G	G	C	16
d	T	T	T	A	C	T	C	A	T	C	G	G	C	3
e	T	T	T	A	C	T	C	A	C	G	G	G	C	7
f	T	T	T	A	C	T	C	A	T	G	A	G	C	1
g	T	T	T	A	C	T	C	A	T	G	G	C	T	4

Site	1	2	3	4	5	6	7	8	9	10	11	12	13	
Root	T	T	T	A	T	C	T	C	T	C	G	G	C	
Lineage														Freq
a	1	1	0	0	0	0	0	0	0	0	0	0	0	25
b	1	1	1	0	0	0	0	0	0	0	0	0	0	1
c	0	0	0	1	0	0	0	0	0	0	0	0	0	16
d	0	0	0	0	1	1	1	1	0	0	0	0	0	3
e	0	0	0	0	1	1	1	1	1	1	0	0	0	7
f	0	0	0	0	1	1	1	1	0	1	1	0	0	1
g	0	0	0	0	1	1	1	1	0	1	0	1	1	4

Site	1	2	3	4	5	6	7	8	9	10	11	12	13	
Root	T	T	T	A	T	C	T	C	T	C	G	G	C	
Lineage														Freq
a	1	1	0	0	0	0	0	0	0	0	0	0	0	25
b	1	1	1	0	0	0	0	0	0	0	0	0	0	1
c	0	0	0	1	0	0	0	0	0	0	0	0	0	16
d	0	0	0	0	1	1	1	1	0	0	0	0	0	3
e	0	0	0	0	1	1	1	1	1	1	0	0	0	7
f	0	0	0	0	1	1	1	1	0	1	1	0	0	1
g	0	0	0	0	1	1	1	1	0	1	0	1	1	4

Sorted columns, small to large
Tree paths are read left to right

	5 6 7 8 13	11	9	10	5 6 7 8	4	3	1 2

	12 13	11	9	10	5 6 7 8	4	3	1 2
a	0	0	0	0	0	0	0	1
b	0	0	0	0	0	0	1	1
c	0	0	0	0	0	1	0	0
d	0	0	0	0	1	0	0	0
e	0	0	1	1	1	0	0	0
f	0	1	0	1	1	0	0	0
g	1	0	0	1	1	0	0	0

Probability distribution of gene trees $(T, \mathbf{n})$

```
25 :  1   2   0

 1 :  3   1   2   0

16 :  4   0

 3 :  5   6   7   8   0

 7 :  9  10   5   6   7   8   0

 1 : 11  10   5   6   7   8   0

 4 : 12  13  10   5   6   7   8   0
```

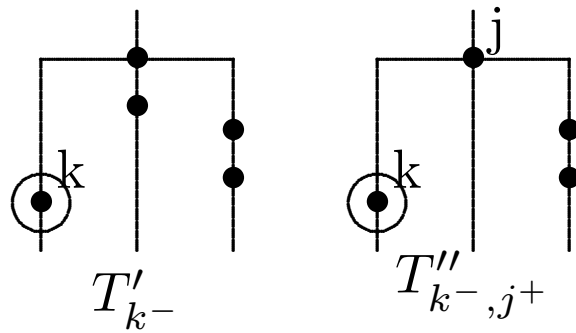Gene tree T contains sequences of mutation paths from the leaves to the root.
$\mathbf{n}$ contains sequence multiplicities.
$p(T, \mathbf{n})$ is the probability of a gene tree.

Probability of a gene tree

$$p(T, \mathbf{n})$$

$$= \frac{(n-1)}{(n-1+\theta)} \sum_{k:n_k \geq 2} \frac{(n_k - 1)}{n - 1} p(T, \mathbf{n} - \mathbf{e}_k)$$

$$+ \frac{\theta}{(n-1+\theta)} \sum_{k} \frac{1}{n} p(T'_{k^-}, \mathbf{n})$$

$$+ \frac{\theta}{(n-1+\theta)} \sum_{k \rightarrow j} \frac{(n_j + 1)}{n} p(T''_{k^-, j^+}, \mathbf{n}'')$$

Removing a mutation.

$$T'_{k^-} \qquad T''_{k^-, j^+}$$

The system is recursive in the degree of $(T, \mathbf{n})$, defined as $n +$ the number of vertices.

Let H be the current configuration $(T, \mathbf{n})$ and H' the configuration immediately prior to the first event back in time corresponding to coalescence or mutation. H' has the form $(T, \mathbf{n} - \mathbf{e}_k)$, $(T'_{k-}, \mathbf{n})$, or $(T''_{k-,j+}, \mathbf{n}'')$. Let C denote the event that the last event was a coalescence, and M that it was a mutation. The recursion for $(T, \mathbf{n})$ is derived by considering (H', C), (H', M).

$$
\begin{aligned}
P(H) &= \sum_{H'} P(H \mid H', C) P(H', C) \\
&\quad + \sum_{H'} P(H \mid H', M) P(H', M) \\
&= \frac{n-1}{n+\theta-1} P(H \mid H', C) P(H') \\
&\quad + \frac{\theta}{n+\theta-1} \sum_{H'} P(H \mid H', M) P(H')
\end{aligned}
$$

In the recursion for $p(T, \mathbf{n})$, $(n_k - 1)/(n-1)$ is the probability that a type k gene with $n_k - 1$ copies while there were $n - 1$ edges in the coalescent tree branched to n_k genes while $n - 1$ edges. In the term where a mutation occurred last, $1/n$ and $(n_j + 1)/n$ are the probabilities that in a coalescent tree of n the mutation falls on an edge which will give the correct current configuration $(T, \mathbf{n})$ from the prior configuration H'.

Ancestral Inference from Gene trees

Coalescent History Process

H_{-m}

$H_{\ell-1}$
H_{ℓ}

H_{k-1}
H_k

H_0
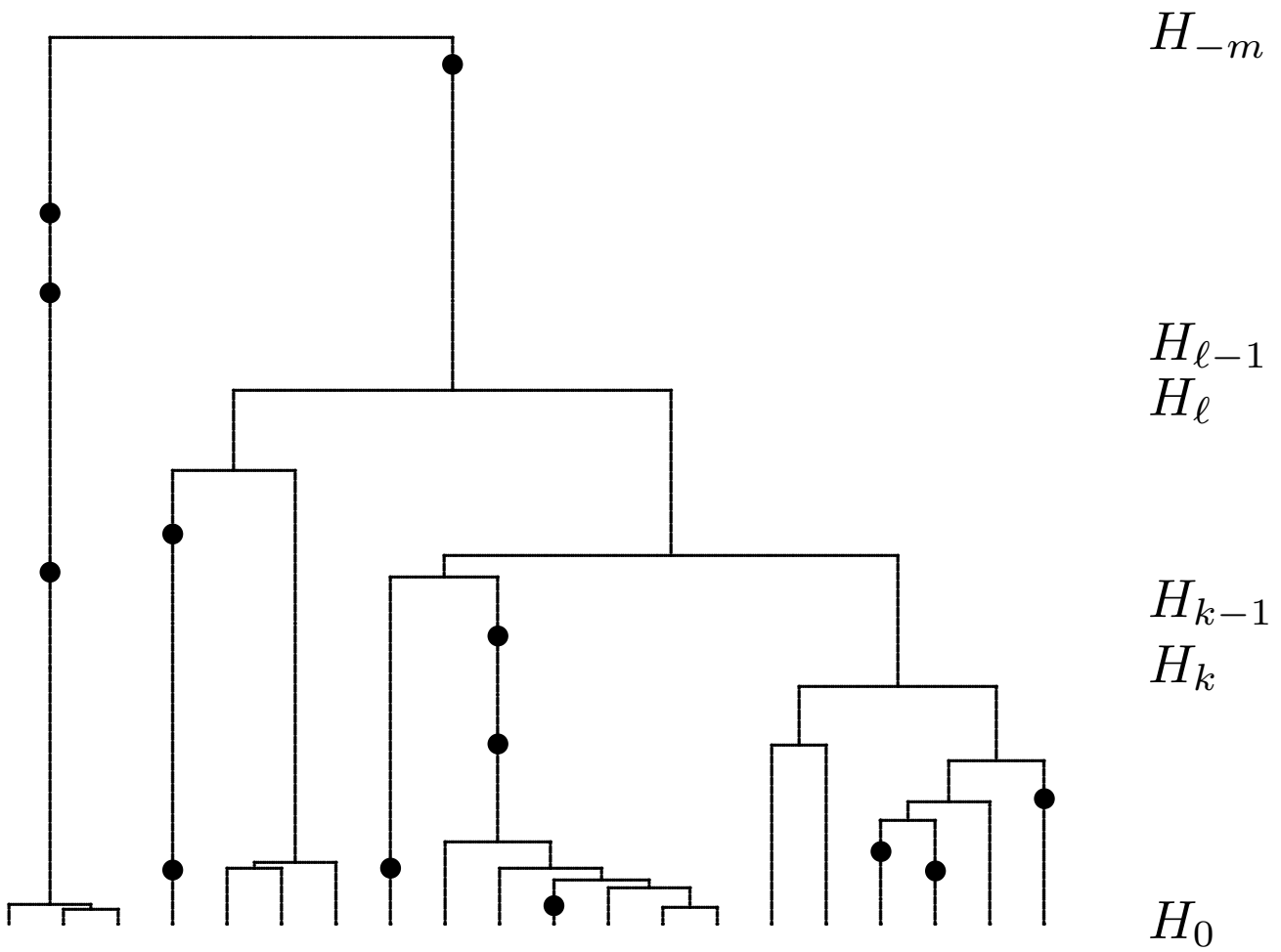
History states H_k are Gene trees $(T, \mathbf{n})$
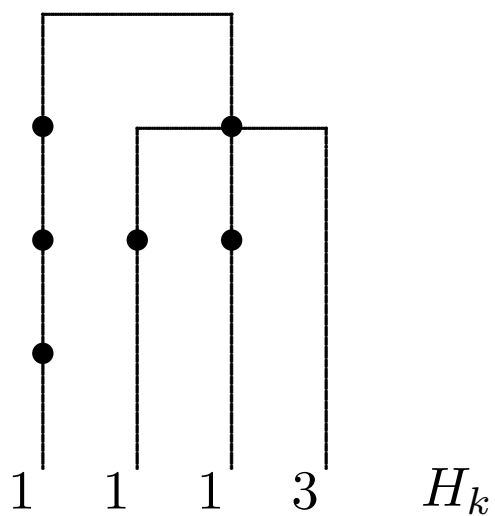
Tree T. Multiplicity of lineages $\mathbf{n} = (1, 1, 1, 3)$.

Sequential Importance Sampling

Let H_j be the history configuration of gene types at step j back in the coalescent process of the sample, where at each step either a mutation or coalescence has occurred back in time.
$H_{-m}, H_{-m+1}, \ldots, H_1, H_0$ is the history process of the sample. A single MRCA is reached at $-m$.

$$p(H_j) = \sum p(H_j \mid H'_{j-1}) p(H'_{j-1}) \;\Big\downarrow$$

Summation is over possible configurations H'_{j-1}. Forward transition probabilities $\downarrow p(H_j \mid H'_{j-1})$ are known.
$p(H_j)$ and $\{p(H'_{j-1})\}$ are unknown.

Reverse IS transition probabilities $\uparrow \hat{p}(H'_{j-1} \mid H_j)$

$$p(H_j)$$

$$= \sum \frac{p(H_j \mid H'_{j-1})}{\hat{p}(H'_{j-1} \mid H_j)} \hat{p}(H'_{j-1} \mid H_j) p(H'_{j-1}) \ \uparrow$$

The importance sampling representation is

$$p(H_0)$$

$$= E_{\hat{p}} \left[\frac{p(H_0 \mid H_{-1})}{\hat{p}(H_{-1} \mid H_0)} \cdots \frac{p(H_{-m+1} \mid H_{-m})}{\hat{p}(H_{-m} \mid H_{-m+1})} \right]$$

$$= E_{\hat{p}} \left[\frac{p(H_0 \mid H_{-1}) \dots p(H_{-m+1} \mid H_{-m}))}{\hat{p}(H_{-m} \mid H_{-m+1}) \dots \hat{p}(H_{-1} \mid H_0)} \right]$$

$$= E_{\hat{p}} \left[\frac{p(H_\downarrow)}{\hat{p}(H_\uparrow)/p(H_0)} \right]$$

The MRCA state H_{-m} is the single root of the tree, and $p(H_{-m}) = 1$.
Simulate $\uparrow$ under $\hat{p}$ repeatedly and average IS weights on histories to obtain the likelihood of the data $p(H_0)$.

Proposal distribution for gene trees $\hat{p}(H_{j-1} \mid H_j)$.

Choose a gene in H_{j-1} uniformly from the possible genes which may change by coalescence or mutation.
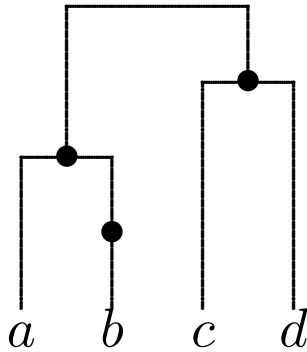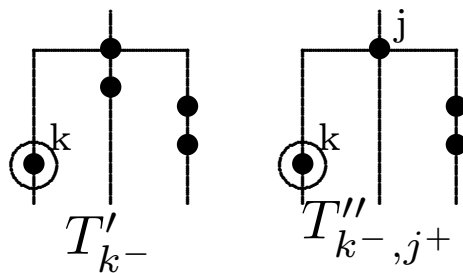

Choose from b, c, d each with probability $1/3$.

(i) b: Remove the mutation on lineage b.
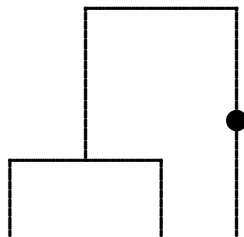
(ii) c or d: Coalesce the two lines c and d.

H_{i-1}	Proposal	Weights
$(T, \mathbf{n} - \mathbf{e}_k)$	$\dfrac{n_k}{n_\circ}$	$\dfrac{n_\circ}{n_k} \cdot \dfrac{n_k - 1}{n - 1 + \theta}$
$(T'_{k^-}, \mathbf{n})$	$\dfrac{1}{n_\circ}$	$\dfrac{n_\circ}{n} \cdot \dfrac{\theta}{n - 1 + \theta}$
$(T''_{k^-,j^+}, \mathbf{n} + \mathbf{e}_j)$	$\dfrac{1}{n_\circ}$	$\dfrac{n_\circ}{n} \cdot \dfrac{(n_j + 1)\theta}{n - 1 + \theta}$

The time spent in a configuration $(T, \mathbf{n})$ is exponential with rate $\frac{1}{2}(n(n-1) + \theta s)$, where s is the number of singleton mutations that are possibilities for removal in the next event back in time.

Example

A simple example is computing the probability of the tree above by importance sampling.

In the first transition back in time either a coalescence takes place or a mutation is removed. $n_0 = 3$.

H_{-1}	Prob	Weights
Coalescence	$\frac{2}{3}$	$\frac{3}{2(2+\theta)}$
Mutation	$\frac{1}{3}$	$\frac{\theta}{2+\theta}$

If coalescence took place then the next transition must remove the mutation with probability 1, and weight $\theta/(1+\theta)$ and then finally coalescence must occur with probability 1, and weight $1/(1+\theta)$.

If a mutation was removed then the next two transitions must be coalescences with probabilities 1,1 and weights $2/(2+\theta)$ and $1/(1+\theta)$.

The algorithm is seen to be generating two types of histories, given the topology, with probabilities and weights

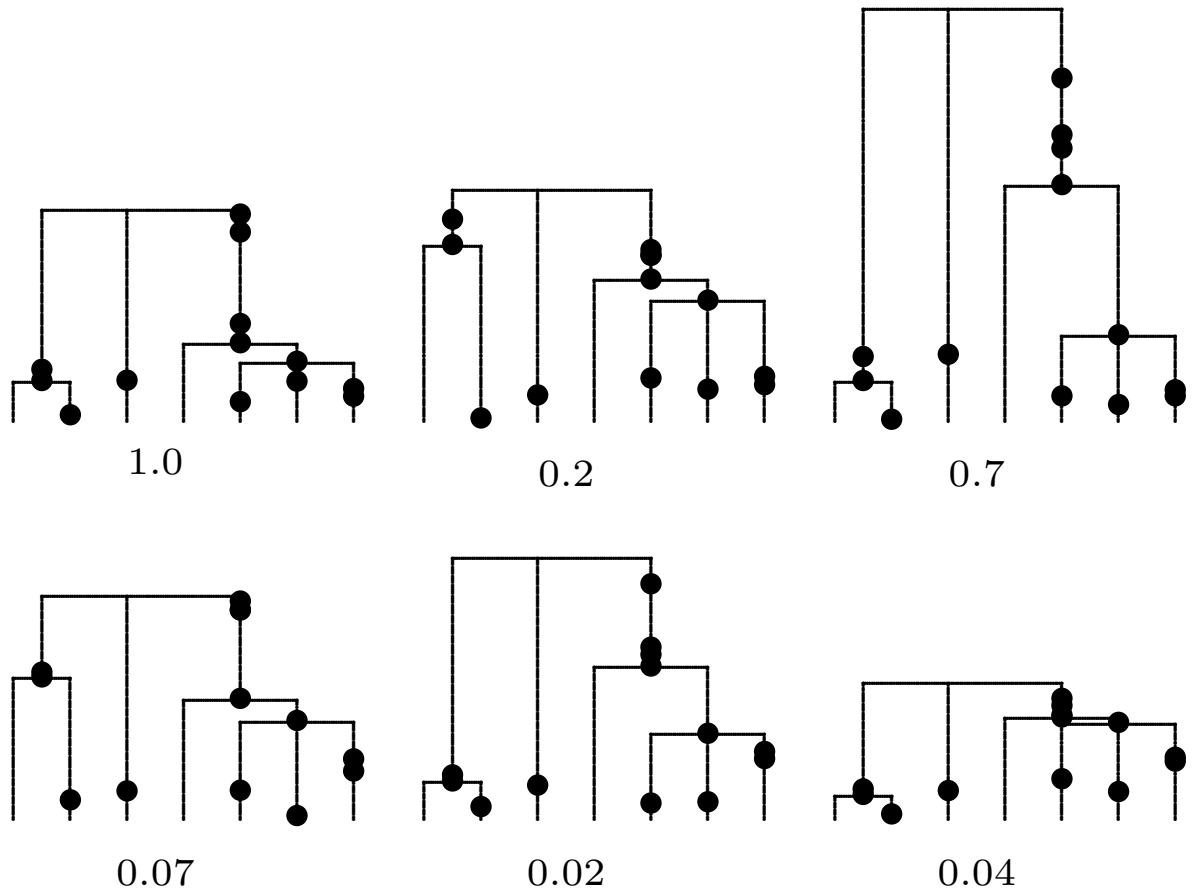| Prob	Weights
$\frac{2}{3}$	$\frac{3\theta}{2(2+\theta)(1+\theta)^2}$
$\frac{1}{3}$	$\frac{6\theta}{(2+\theta)^2(1+\theta)}$

This shows that the algorithm will (correctly) generate

$$L = \frac{\theta}{(2+\theta)(1+\theta)^2} + \frac{2\theta}{(2+\theta)^2(1+\theta)}$$

for the probability of the genetree, and

$$\frac{\theta}{(2+\theta)(1+\theta)^2 L}, \quad \frac{2\theta}{(2+\theta)^2(1+\theta)L}$$

for the two different histories, conditional on the topology.

Simulated gene trees with relative likelihoods conditional on tree topology

1.0

0.2

0.7

0.07

0.02

0.04

TMRCA and ages of mutations

Importance sampling simulates coalescent histories back in time that are compatable with the topology of the gene tree. Each simulation run gives a likelihood value l, and a history. If the likelihood values returned in r runs are $l_1, \ldots, l_r$ then an estimate of the likelihood is the mean $\bar{l}$. An empirical distribution of the TMRCA is $(t_1, p_1), \ldots, (t_r, p_r)$ where the t_j are simulated TMRCA values, and $p_j = l_j / \bar{l}$. The betaglobin tree introduced in the first lecture is a gene tree drawn to scale with mean ages and mean TMRCA from their empirical distributions. Software `genetree` is available to compute the mean ages of mutations and TMRCA.

Griffiths, R.C. and Tavaré S. (1999). *The ages of mutations in gene trees.* Ann. Appl. Prob.**9**, 567–590.

Griffiths, R. C. (2002). Ancestral inference from gene trees. In: Veuille, M. and Slatkin, M. (Ed.), *Modern Developments in Theoretical Population Genetics: the Legacy of Gustave Malcot*, Oxford University Press, New York, pp. 94–117.

Stephens, M., & Donnelly, P. (2000). Inference in molecular population genetics. *Journal of the Royal Statistical Society: Series B* 62, 605–55.

Tavaré S. (2004). *Ancestral inference in population genetics.* In: Picard, J. (Ed.), *Lectures on Probabiity and Statistics.* Ecole d'Eté de Probabilités de Saint-Flour XXXI- 2001, Springer, Berlin, pp. 1–188.