The Coalescent

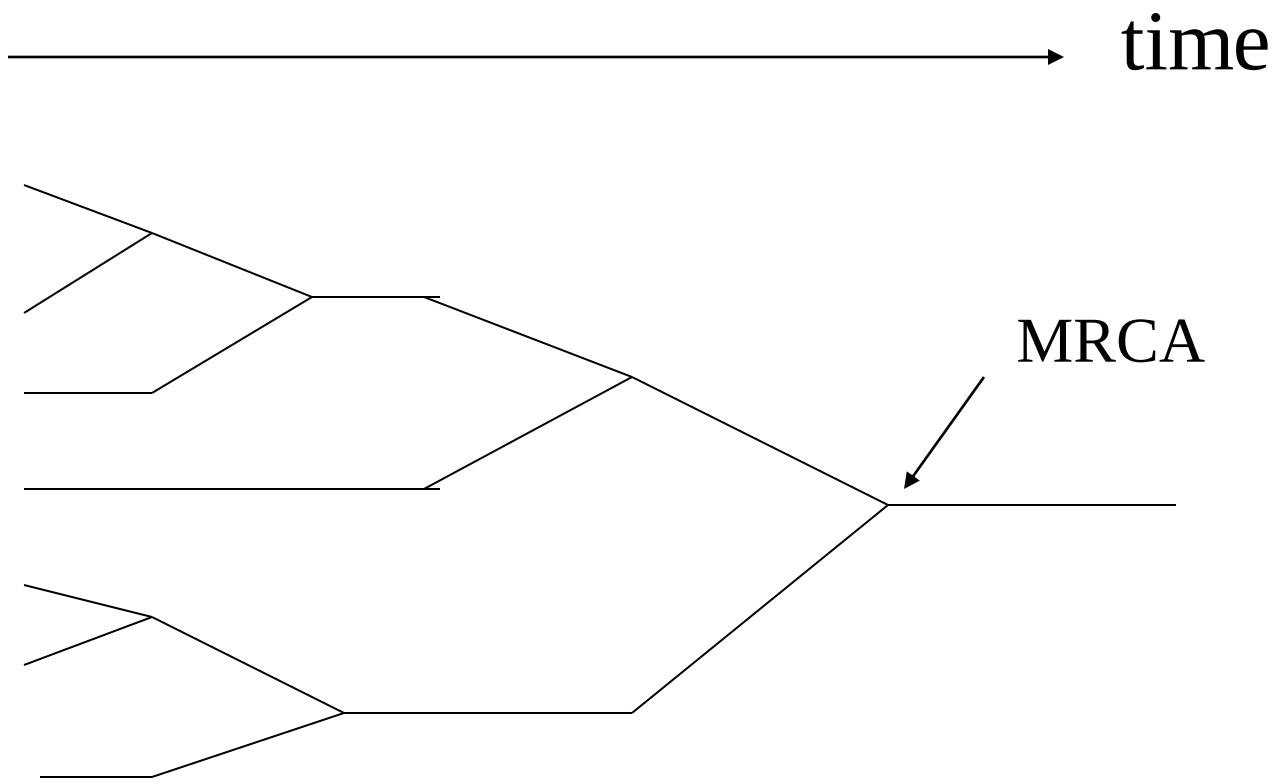
The Coalescent

- We examined Finite Populations via the notion of IBD. If we wish to simulate such a population (on the computer) then we would need to take N haploid individuals and generate their N offspring, with appropriate alleles for a succession of generations. This is very heavy computationally.

The Coalescent

- An alternate approach which has great utility for some situations, is to study the coalescent process. We look at the process back through time.
- Start with n individuals and study the process defined by their ancestors in each successive generation (working backwards)

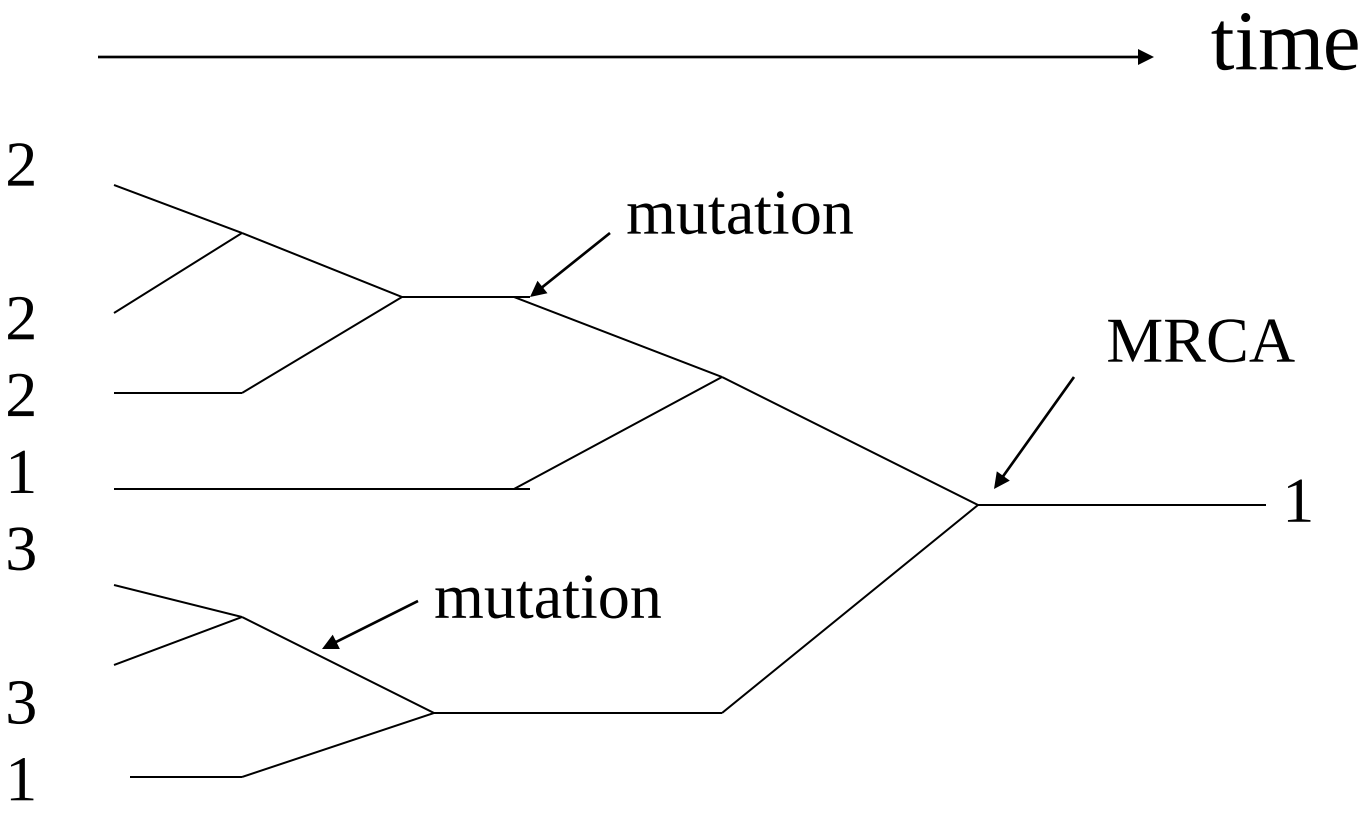
The Coalescent



The Coalescent

- Now we can add mutations into the process, supposing that we know the length of the edges (times)

The Coalescent

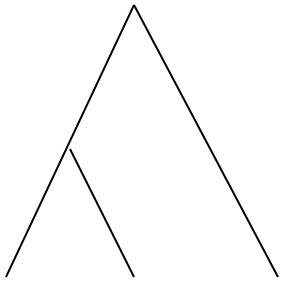


The Coalescent

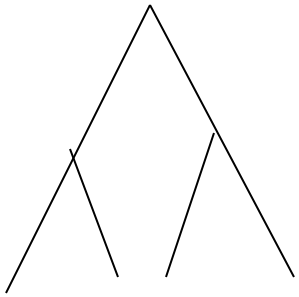
- Essentially we have separated the genealogical process (the tree) from the mutation process.

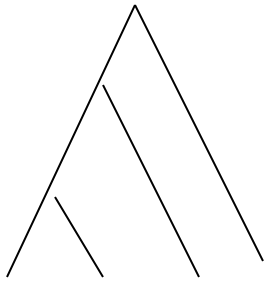
The Coalescent

- Trees.
- n=3



- n=4

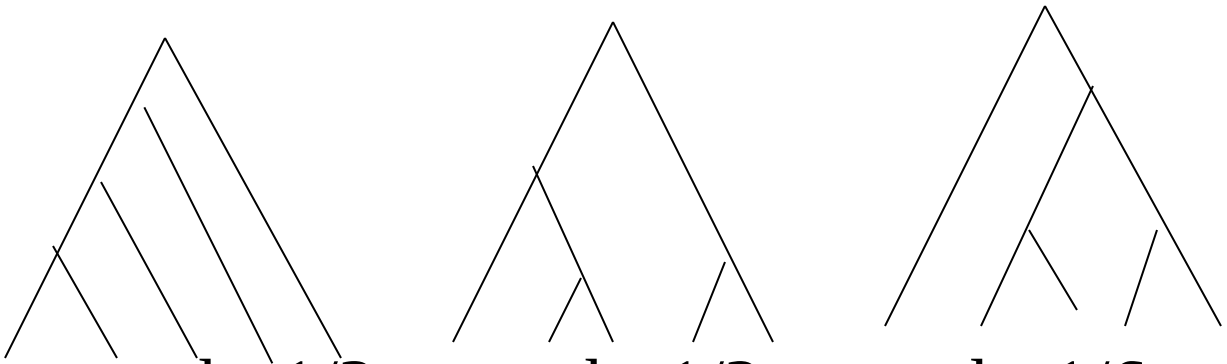




prob=1/3 prob=2/3

The Coalescent

n=5



prob=1/3 prob=1/2 prob=1/6

Distribution of Tree Topology

Top-down (for n leaves)

First sub-trees have r and (n-r) leaves with probabilities 1/(n-1), then their subtrees split similarly, but you need to allow for the choice of next subtree to split.

Distribution of Tree Topology

Thus for n=5,

(4,1),(3,2),(2,3),(1,4) then

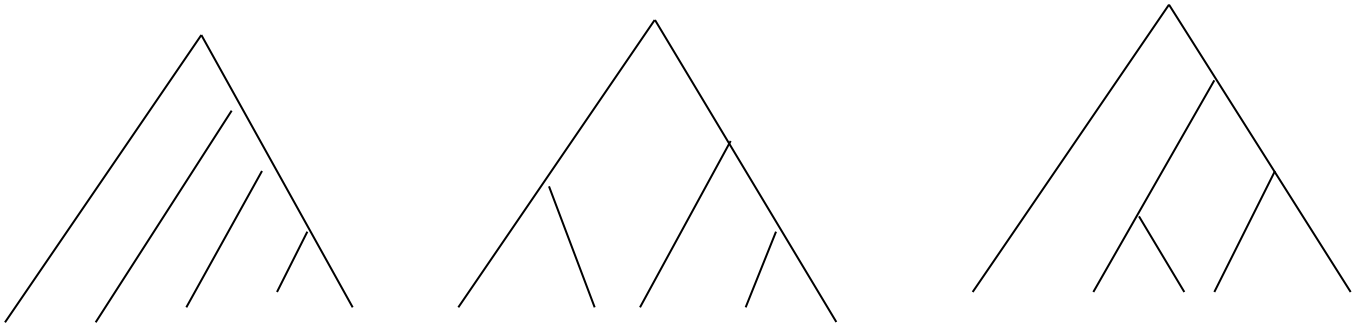
((3,1),1),((1,3),1),

((2,1),2),(3,(1,1)) etc

((((1,1),1),1),1),(((1,(1,1),1),1),((1,((1,1),1),1),
((1,(1,(1,1),1),1)
etc

Distribution of Tree Topology

- n=5



8 12 4

The Coalescent

- Now suppose that for a population of N individuals we use the Wright-Fisher model. The assumption of this model is that each offspring “selects” its parent at random from the N in the previous generation.
- Consider 2 individuals in generation 0 (time will be measured backwards) and their lineages back through time.

The Coalescent

- These two lineages will at some time coalesce. When? The probability that they coalesce at time 1 is 1/N, i.e. Prob{Not coalesced by t=1}=(1-1/N).
- Each step back is independent so Prob{Not coalesced by t}=$(1-1/N)^t$

$=(1-1/N)^{N\tau}$

(where $\tau=t/N$)

approx=$e^{-\tau}$

The Coalescent

- Consider Lim $(1-a/N)^N$ as N tends to inf.

$$(1-a/N)^N = (1 - (a/N)N + (a/N)^2N(N-1)/2! - (a/N)^3N(N-1)(N-2)/3! + \ldots\ldots$$

$$= 1 + (-a) + (-a)^2/2! + (-a^3)/3! + + (-a)^r/r! + \ldots\ldots + o(1/n)$$

$$= e^{-a} + o(1/N)$$

The Coalescent

- Thus the
  Prob{2 lineages not coalesced by τ}=$e^{-\tau}$ approx, which is a negative exponential with rate 1, and the expected time to coalescence is 1 unit of time (=N generations).

The Coalescent

- Now consider k lineages (perhaps corresponding to a sample from the current generation)
  Prob{none coalesce in the previous gen}=
  Prob{they have k parents}=
  $(1-1/N)(1-2/N)...(1-(k-1)/N)=_{i=0}\Pi_{i=k-1}(1-i/N)$
  approx = $1 - {}_kC_2/N + O(1/N^2)$.

The Coalescent

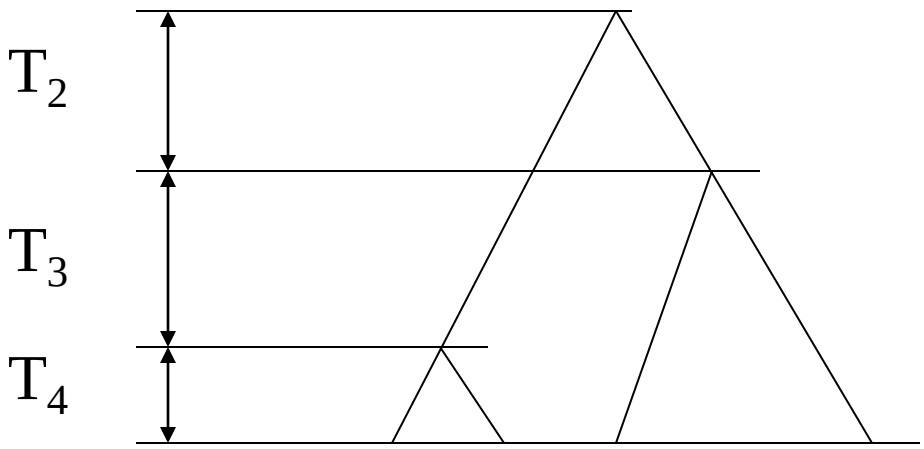
- Now the probability that more than two coalesce in any generation is negligible.
- Let T(k)=(scaled) time to first coalescence then, as earlier, T(k) is neg. exp. with mean 2/(k(k-1)), and these coalescences take place one at a time.

The Coalescent.MRCA

- Thus if we start with k we have T(k), T(k-1) ,..., T(3),T(2) as the times back to the MRCA (Most Recent Common Ancestor).
- Take $TT(k) = \sum_{i=2}^{i=k} T(i)$, so the expected time $E(TT(k)) = \sum_{i=2}^{i=k} 2/(i(i-1))$
  $= \sum_{i=2}^{i=k} 2\,(1/(i-1) - 1/i)$
  $= 2(1-1/k)$ approx $= 2$ for large k

The Coalescent.MRCA

- Thus E(TT(2)) = 1 is more than ½ of the expected time for the tree for any k.
- Typical tree n=4



$E(T_2) = 1$
$E(T_3) = 1/3$
$E(T_4) = 1/6$

The Coalescent.

- Variances

$V(T_i) = (2/(i(i-1)))^2$, $V(T_2) = 1$,

$TT(k) = {}_{i=2}\Sigma_{i=k}\, T(i)$, so

$V(TT(k)) = {}_{i=2}\Sigma_{i=k}\, (2/(i(i-1)))^2$

$= {}_{i=2}\Sigma_{i=k}\, 2^2\, (1/(i-1) - 1/i)^2$

$= 4\ {}_{i=2}\Sigma_{i=k}\ (1/(i-1)^2 + 1/i^2 - 2/(i(i-1)))$

The Coalescent.

- $= 4 \sum_{i=2}^{i=k} (1/(i-1)^2 + 1/i^2 - 2/(i(i-1)))$

  $= 8 \sum_{i=1}^{1} 1/i^2 + 4/k^2 - 4 - 8(1-1/k)$ from earlier

  $= 8 \sum_{i=1}^{i=k-1} 1/i^2 - (3k+1)(k-1)/k^2$

The Coalescent. V(TT(k))

- $V(TT(k)) <= 8\pi^2/6 - 12$ approx = 1.16 so that $V(T_2)$ accounts for most of the variance.
- The following shows six realisations for a sample of size five.

The Coalescent.

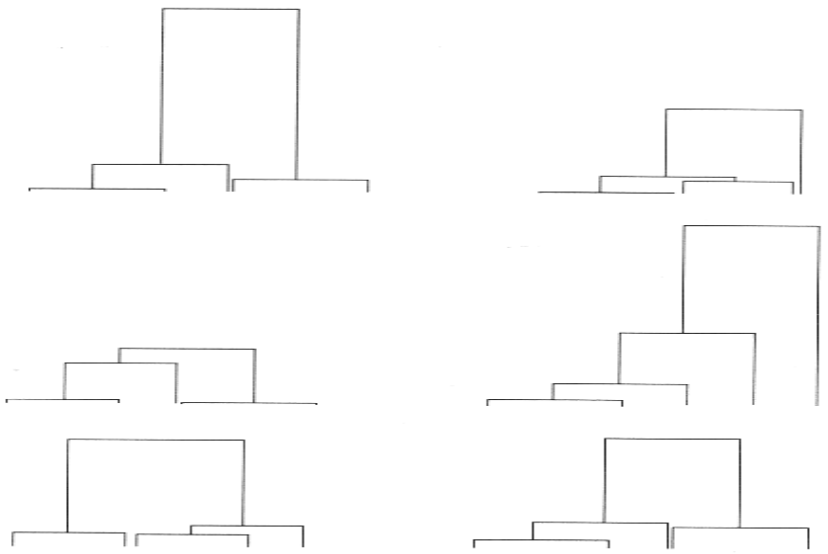


Fig. 2.3. Six realizations, drawn on the same scale, of coalescent trees for a sample of $n = 5$. (In each tree the labels 1,2,3,4,5 should be assigned at random to the leaves.)

© S. Tavaré and P. Donnelly. DRAFT of September 15, 1999
Duplication for commercial purposes prohibited

The Coalescent.

- Adding recombination to the forward process on the tree is somewhat difficult, and requires MCMC (Monte Carlo Markov Chain) a numerical estimation method.