The Coalescent

Chris Cannings,
University of Sheffield,
c.cannings@shef.ac.uk
http://www.amorph.group.shef.ac.uk/

Singapore, 19.05.2011

The Coalescent

(1) Genetic Drift. Stochastic model of evolution of population of neutral genes.

(2) The Coalescent. Reversing time.

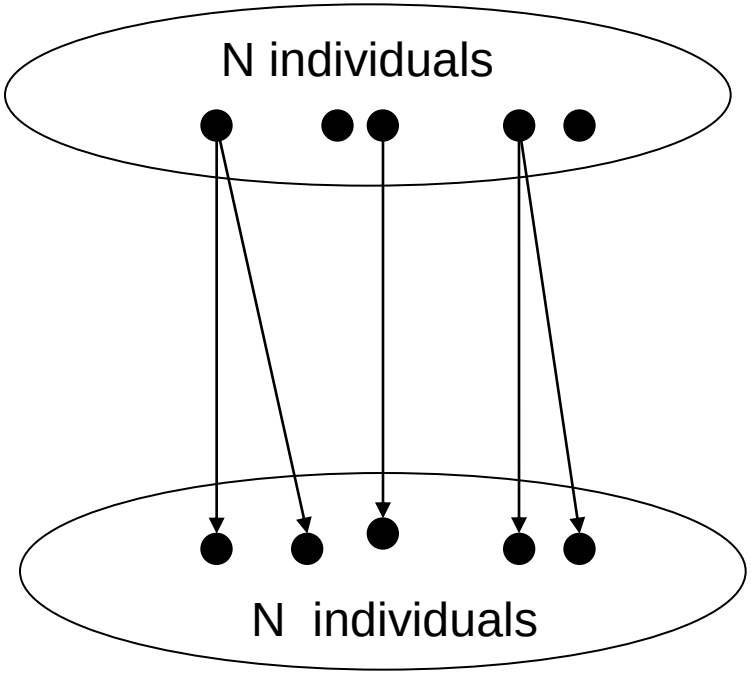
(3) Tree topology. Catalan numbers.

(4) External Edges. Distribution of lengths.

(5) Stepwise Mutation Model. Distributions and moments. Bell Polynomials.

Genetic Drift

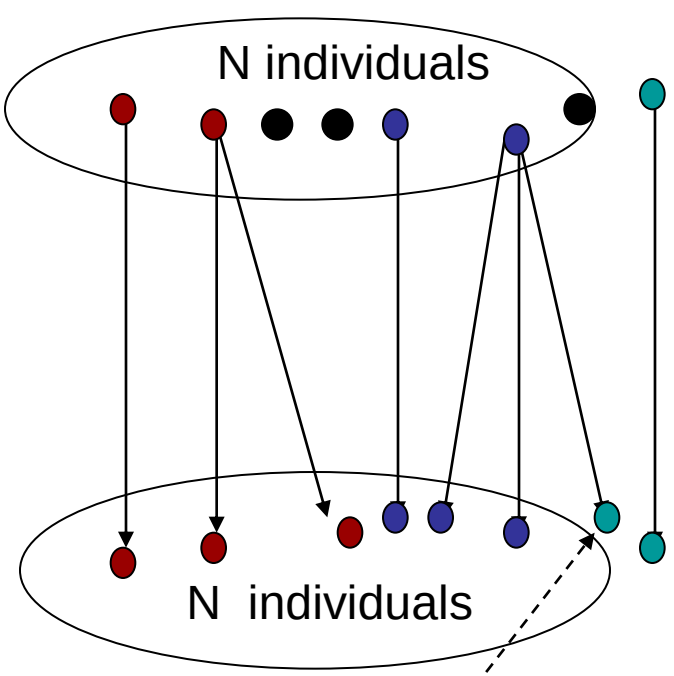
Unisexual



discrete
generations

Genetic Drift

Individuals of varying types. Possibly mutation, Recombination, etc.



discrete generations

Mutation (offspring type different from parent type)

Wright-Fisher Model

- Each individual in the m’th generation is the offspring of an individual in the (m-1)’th generation selected at random (prob=1/N) independently of all others.

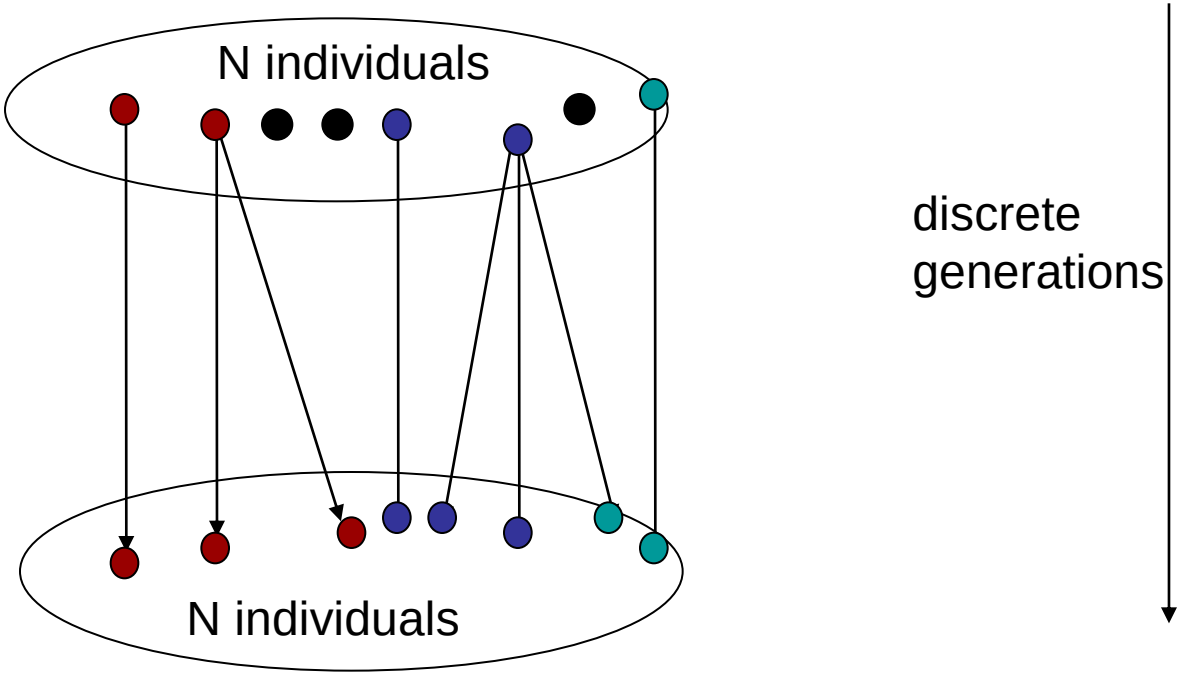
Moran Model

- One individual dies and one (possibly the same one) gives birth.

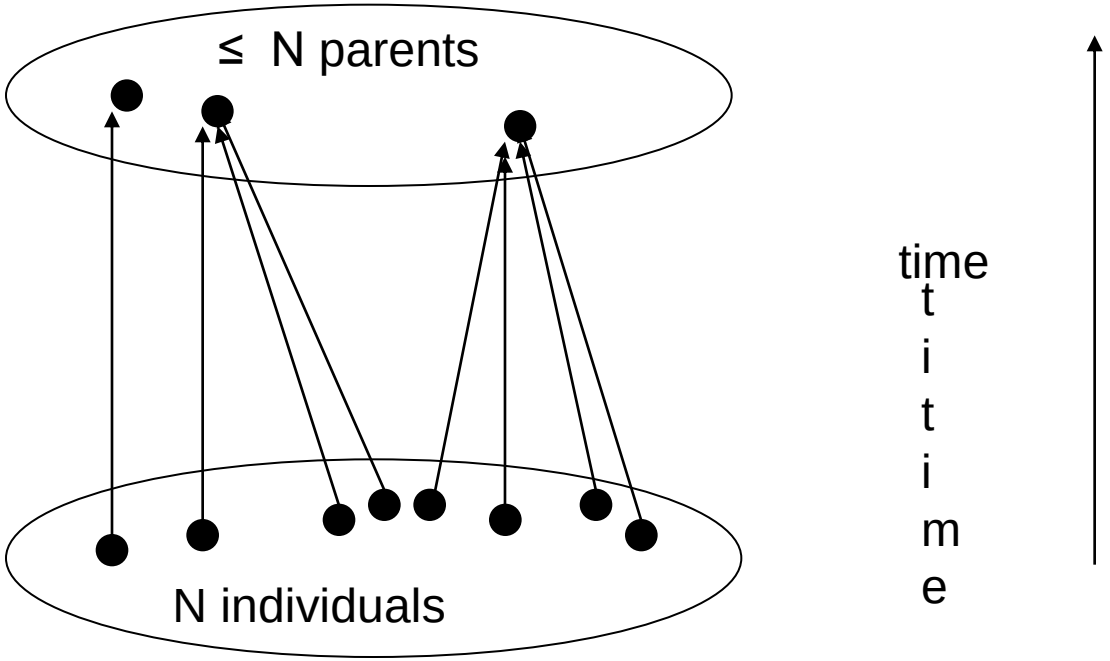
Canning's Model

- Cannings(1974) the n individuals produce $X_1, X_2, \ldots\ldots\ldots, X_n$ offspring these being exchangeable r.v.'s. Markov Chain
- This model (which encompasses the Wright-Fisher & Moran model) gives rise to neat formulae for the eigenvalues and eigenvectors of the process.

Genetic Drift (multiple types)



The Coalescent

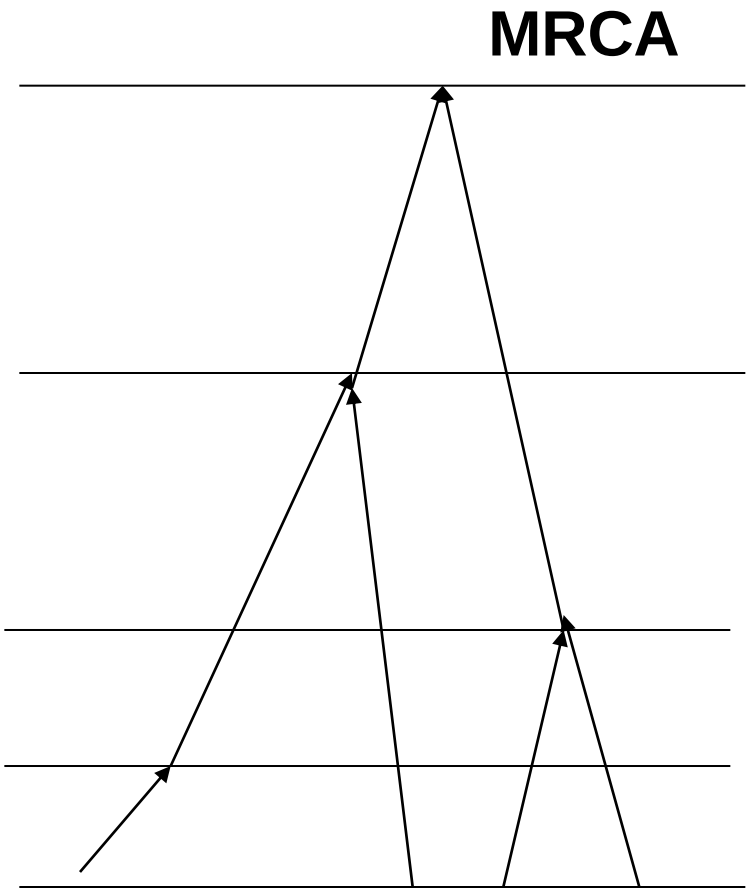


Kingman, 1982

Coalescent

- For a finite population undergoing random reproduction (i.e. not just one offspring for each adult) the current generation will all be descended from a single common ancestor (provided the population has been running for sufficient time) the Most Recent Common Ancestor (MRCA).
- Tracing back to that MRCA produces a tree, and in this tree lines coalesce as we run backwards in time.

Coalescent Tree



Coalescent

- The coalescent allows one to track the behaviour of a population without the necessity of keeping track of all N individuals through the generations.
- Process separated into two pieces
  (1) produce the coalescent tree (time backwards)
  (2) run types, with mutation, recombination (time forward).

Coalescent

- At each point in time (which is run backwards) keep track only of the individuals who are ancestors of the n individuals one is interested in (or sampled) at time 0.

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

(where $\tau=t/N$)

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 = $1 - {}_kC_2/N + O(1/N^2)$.

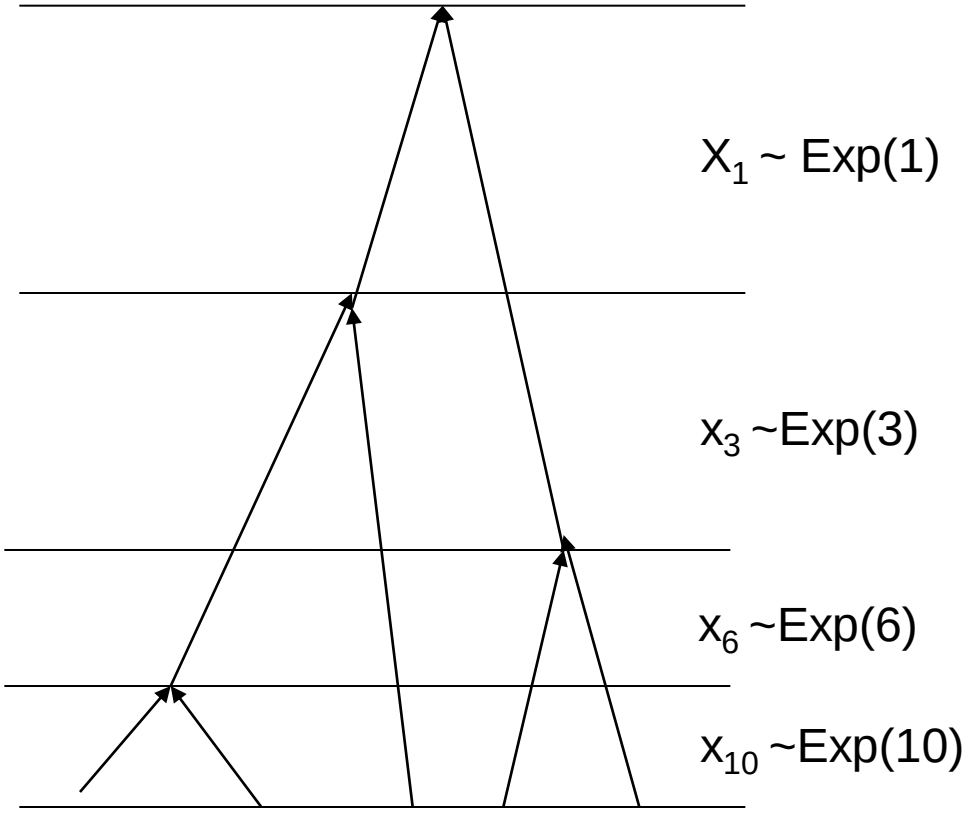
The Coalescent

- 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.

Coalescent (summary)

- At each point in time (which is run backwards) keep track only of the individuals who are ancestors of the n individuals at time 0.
- Replace discrete generations with continuous time, and assume that any pair of individuals at time t will be the offspring of one individual at t-δt with same prob (λδt) as any other pair, independently
- Scale time

The Coalescent (n=5)



Assume no multiple coalescents, only pairwise allowed.

The Coalescent

- The times we are interested in will be sums of negative exponentials (see last slide)
- There is a general expression for a sum of negative exponential random variables.

$\sum_i \eta_i$ for independent $\eta_i \sim Exp(\lambda_i)$

$$\sum_i^l \eta_i \text{ has density function}$$

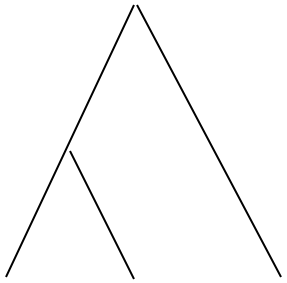
$$\sum_i^l \lambda_i exp(-\lambda_i t)\{\prod_{j \neq i}^l \lambda_j/(\lambda_j - \lambda_i)\}.$$

Tree topology

- We need to find the probabilities for the possible tree topologies (shapes)

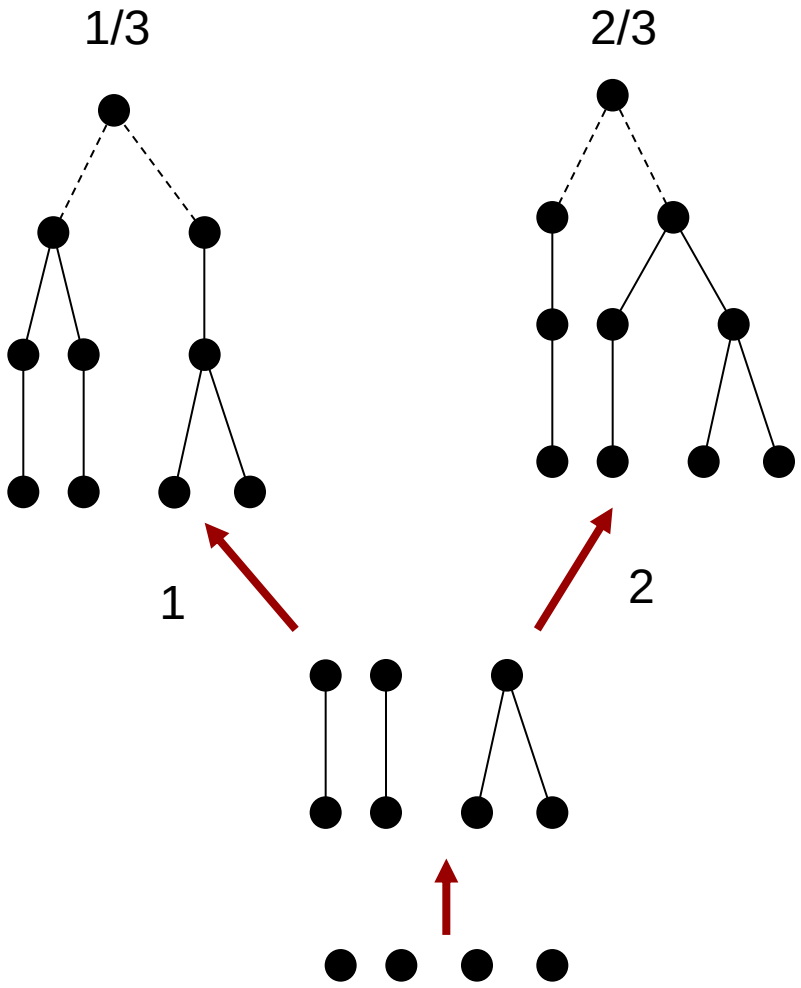
n=3

- n=3



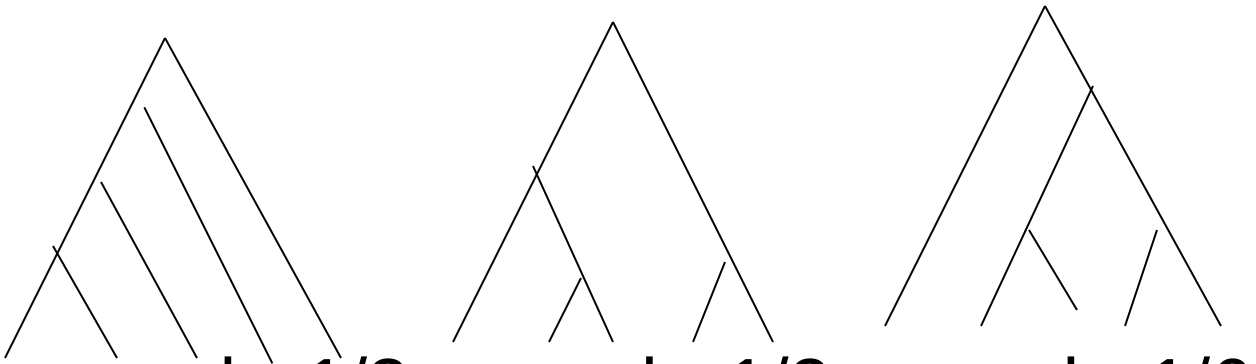
Only one possible shape.

n=4



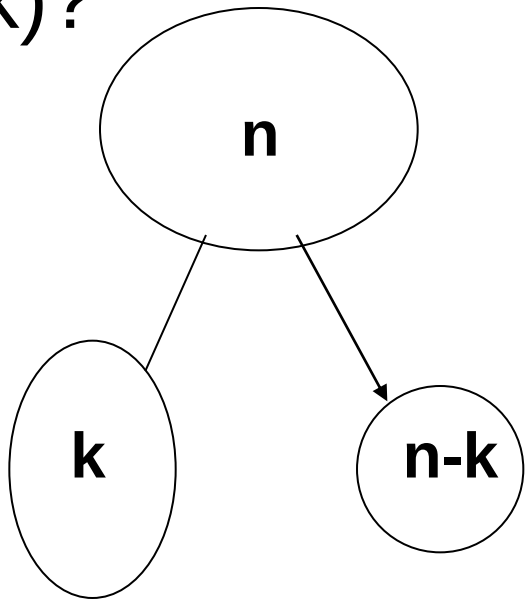
The Coalescent

n=5



Distribution of Tree Topology

- Suppose we have a population of size n; what is probability that we have a split at top of (k,n-k)?



Distribution of Tree Topology

- There will be

$$ {}_nC_{2\ n-1}C_{2\ n-2}C_2 \cdots {}_2C_2 = {}_nC_{2\ n-1}C_2 \ldots\ldots\ldots {}_3C_{2\ 2}C_2 = \frac{n(n-1)}{2}.\frac{(n-1)(n-2)}{2}\ldots\ldots\frac{3.2}{2}.\frac{2.1}{2} = n((n-1)!)^2/2^{n-1} $$

orders for chosing the pairs, and

$ {}_kC_{2\ k-1}C_2 \ldots {}_2C_2 = k((k-1)!)^2/2^{k-1} $ within the k and

$ {}_{n-k}C_{2\ n-k-1}C_2 \ldots\ldots\ldots {}_2C_2 = (n-k)((n-k)!)^2/2^{(n-k)-1} $ within the (n-k).

Distribution of Tree Topology

Now the k-1 joins in the k set and the (n-k-1) in the (n-k) set can be ordered in ${}_{n-2}C_{k-2}$ ways so the probability of k/(n-k) is

$$\frac{{}_nC_k\ {}_{n-k}C_{k-2}\ k((k-1)!)^2/2^k\ (n-k)((n-k)!)^2/2^{n-k}}{n((n-1)!)^2/2^n}$$

$$=1/(n-1)$$

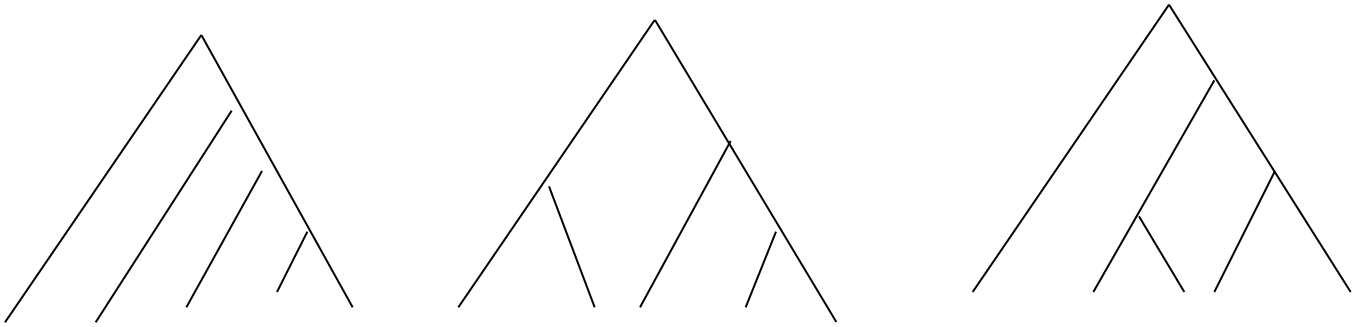
Distribution of Tree Topology

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.

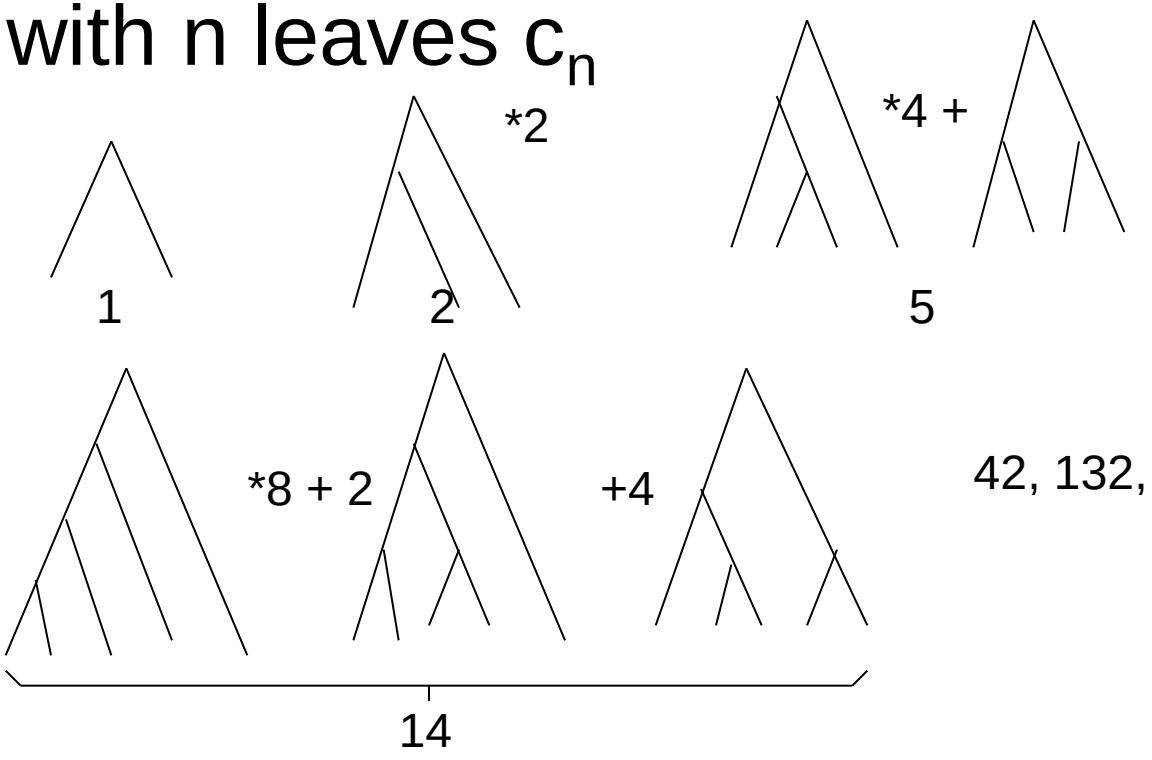
Distribution of Tree Topology

- n=5



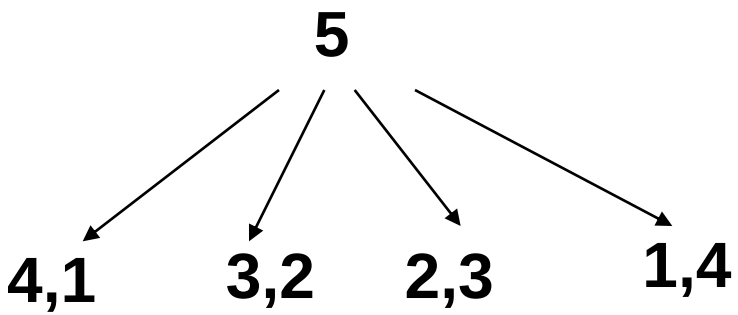
Catalan Numbers (a brief detour)

Number of distinct rooted tree topologies with n leaves c_n



Catalan Numbers (a brief detour)

Example



$$C_n = \sum_{i=1}^{n-1} C_i C_{n-i}$$

Catalan Numbers

- Generating function

$$C(x) = \sum_n c_n x^n$$

Catalan Numbers

- Generating function

$$C(x) = \sum_n c_n x^n$$

now clearly for n>1

$$c_n = \sum_{i=1}^{n-1} c_i c_{n-i}$$

Catalan Numbers

- Generating function

$$C(x) = \sum_n c_n x^n$$

now clearly for n>1

$$c_n = \sum_{i=1}^{n-1} c_i c_{n-i}$$

so

$$[C(x)]^2 = C(x) - x$$

Catalan Numbers

- Generating function

$$C(x) = \sum_n c_n x^n$$

now clearly for n>1

$$c_n = \sum_{i=1}^{n-1} c_i c_{n-i}$$

so

$$[C(x)]^2 = C(x) - x$$

and so

$$\boxed{C(x) = (1 - (1 - 4x)^{1/2})}$$

Catalan Numbers

- Generating function

$$C(x) = \sum_n c_n x^n$$

now clearly for n>1

$$c_n = \sum_{i=1}^{n-1} c_i c_{n-i}$$

so

$$[C(x)]^2 = C(x) - x$$

$$C(x) = (1 - (1 - 4x)^{1/2})$$

and so

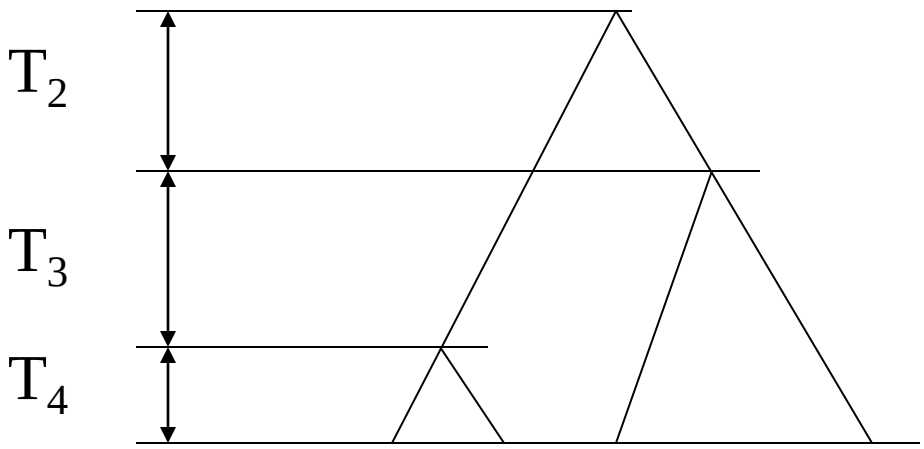
$$c_n = \frac{(2(n+1))!}{n!(n+1)!}$$

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) $= {}_{i=2}\Sigma_{i=k}\, T(i)$, so the expected time $E(TT(k)) = {}_{i=2}\Sigma_{i=k}\, 2/(i(i-1))$
  $= {}_{i=2}\Sigma_{i=k}\, 2\,(1/(i-1) - 1/i)$
  $= 2(1-1/k)$ 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

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

$TT(k) = \sum_{i=2}^{i=k} T(i)$, so

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

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

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

The Coalescent

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

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

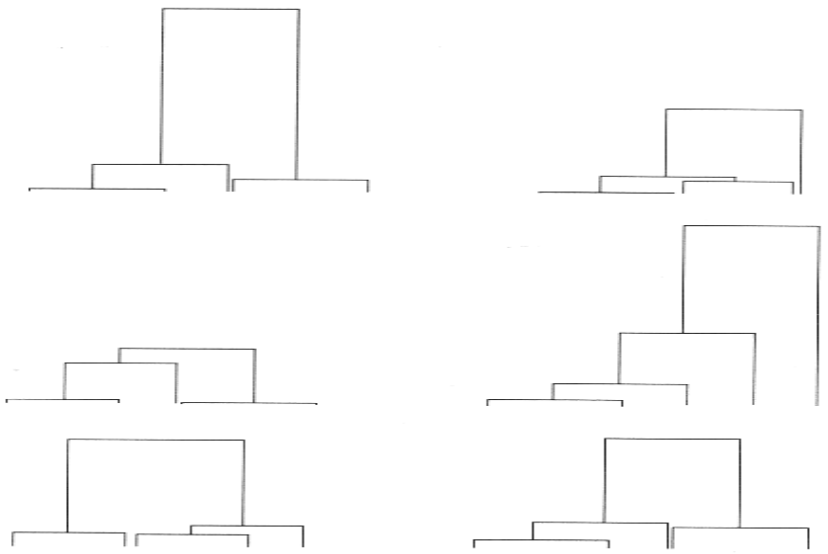
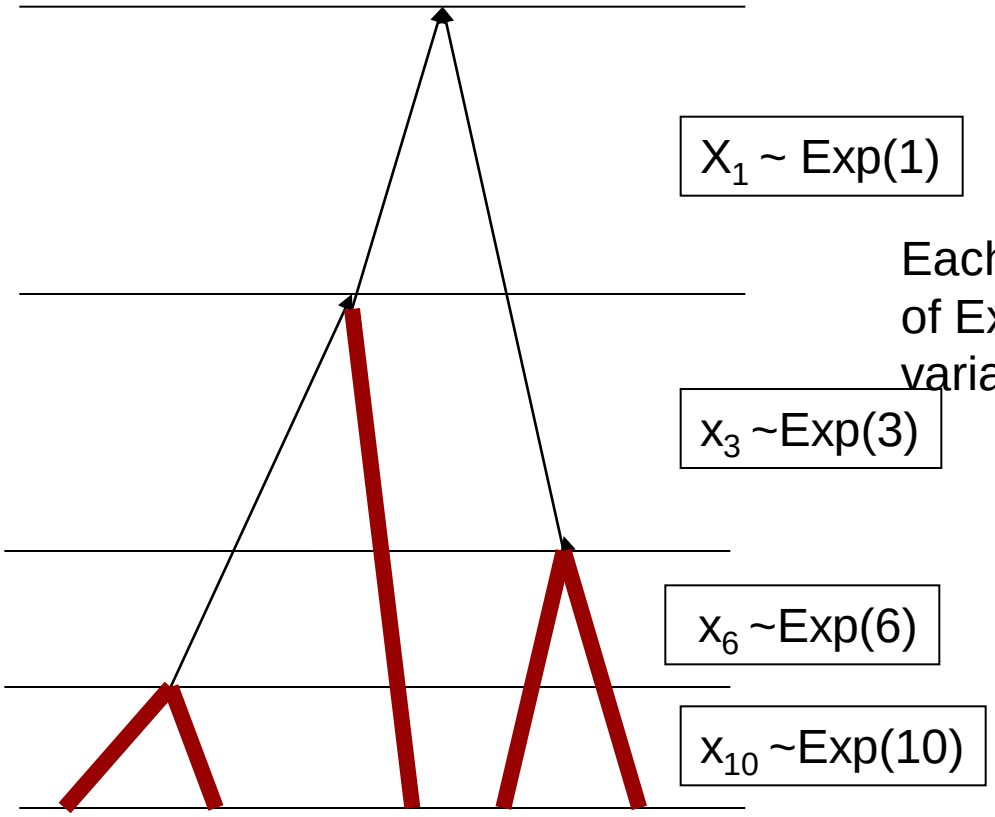


Fig. 2.3. Six realizations, drawn on the same scale, of coalescent trees for a sample of $n = 5$. (In each tree the labels 1,2,3,4,5 should be assigned at random to the leaves.)

The Coalescent.

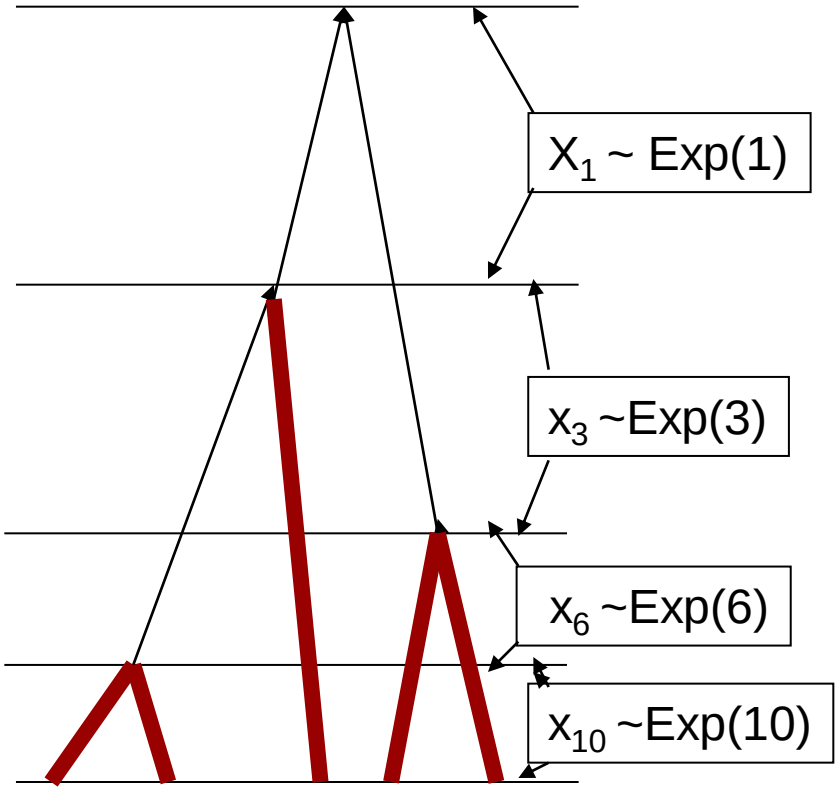
- Adding recombination to the forward process on the tree is somewhat difficult, and requires MCMC (Monte Carlo Markov Chain) a numerical estimation method.

External Edges



Each external edge is the sum of Exponential random variables

External Edges



Here a random edge is x_{10}, $x_{10}+x_{6,,}$ or $x_{10}+x_6+x_3$ with probabilities 2/5, 2/5 and 1/5

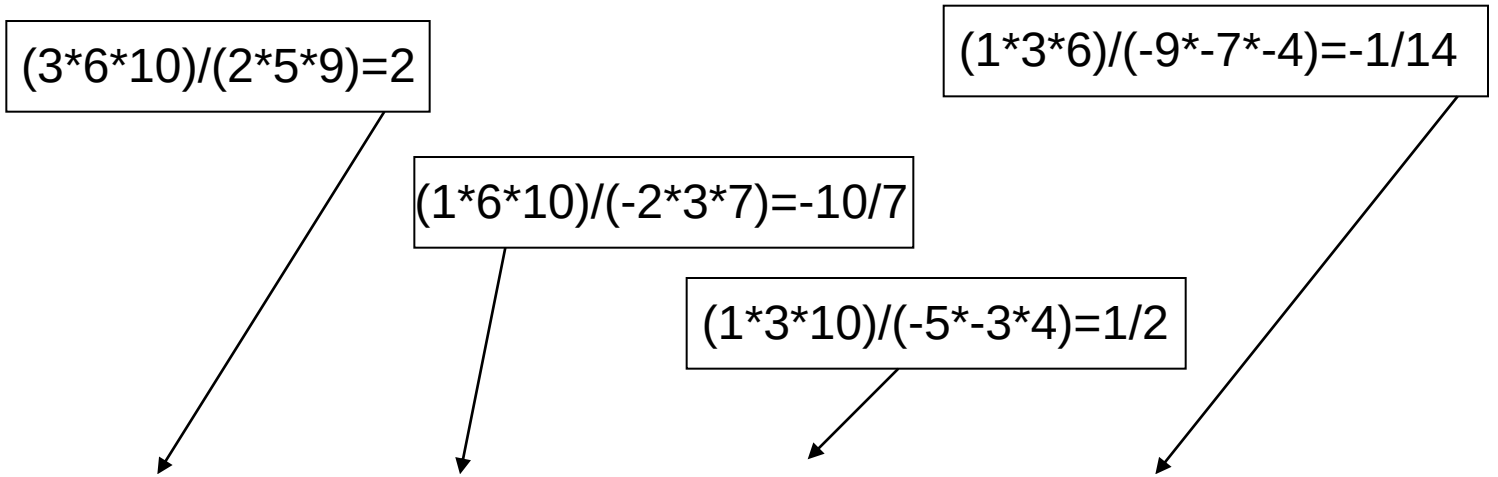
NB. Sums of consecutive $x_{i's}$ from n to k

$\sum_i \eta_i$ for independent $\eta_i \sim Exp(\lambda_i)$

$$\sum_i^l \eta_i \text{ has density function}$$

$$\sum_i^l \lambda_i exp(-\lambda_i t)\{\prod_{j\neq i}^l \lambda_j/(\lambda_j - \lambda_i)\}.$$

$\sum_i \eta_i$ for λ's =1,3,6,10



$$2E(1) - (10/7)E(3) + (1/2)E(6) - (1/14)E(10),$$

where $E(i) = \lambda_i exp(-\lambda_i t)$.

NB. Coefficients alternate in sign.

Prob of trees

- We can look at probability of the different tree topologies.
- Easier to find distribution of edge probabilities.

Probability(edge is Σ n->k)

- Pick one of the n individuals then its external edge is Exp($_nC_2$) with prob=2/n=(n-1) / $_nC_2$.
- Edge is Exp($_nC_2$)+Exp($_{(n-1)}C_2$) with prob=(1-2/n)*2/(n-1)=(n-2) / $_nC_2$.
- Edge is $\Sigma_{i=n,k}$Exp($_iC_2$) with prob (k-1)/ $_nC_2$

n=5

Table 1. n=5

	$\lambda's$	p_k	$a(10exp(-10t))$	$a(6exp(-6t))$	$a(3exp(-3t)$	$exp(-t)$
1	10	4/10	1			
2	10, 6	3/10	−6/4	10/4		
3	10, 6, 3	2/10	9/14	−35/14	40/14	
4	10, 60, 3, 1	1/10	−1/14	7/14	−20/14	28/14
	totals	1	5/70	21/70	30/70	14/70

Distributions n=2(1)6

Table 2

n	e^{-t}	$3e^{-3t}$	$6e^{-6t}$	$10e_{-10t}$	$15e^{-15t}$
2	1				
3	1/2	1/2			
4	3/10	5/10	2/10		
5	14/70	30/70	21/70	5/70	
6	6/42	15/42	14/42	6/42	1/42

NB. All the coefficients are positive.

Result on sequence of exp

- Suppose that we have a set of λ_i's and the coefficient of $\exp(\lambda_k)$ is a_k in the distribution of the random variable which is the sum of the exponential random variables. Now add an extra term with μ then the coefficient of $\exp(\lambda_k)$ is simply $a_k \times \mu/(\lambda_k - \mu)$ irrespective of the details of the sequence of λ_i's.

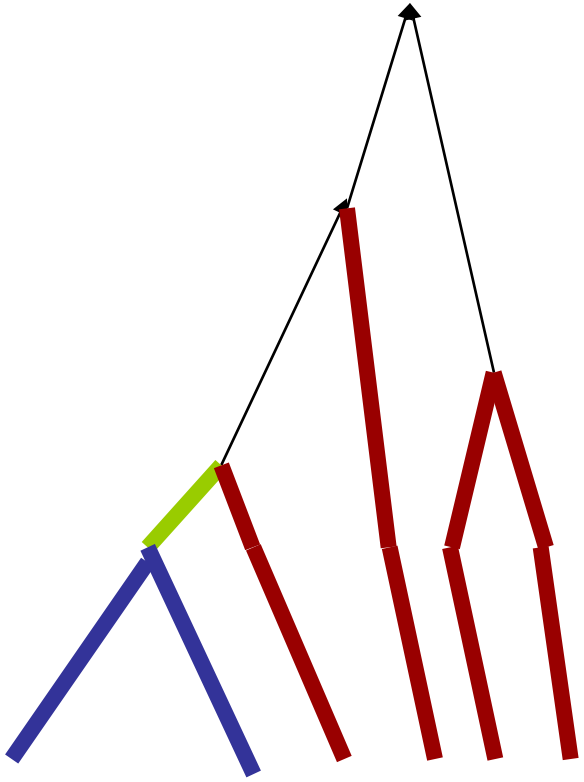
Adding an Extra exp(μ)

$$\sum_i^l \lambda_i exp(-\lambda_i t)\{\prod_{j\neq i}^l \lambda_j/(\lambda_j - \lambda_i)\}.$$

Then when extra exp(μ) term in λ_i changes to

$\lambda_i \exp(-\lambda_i t)\ \Pi\lambda_j/\lambda_j-\lambda_i)\ X\ \mu/(\mu-\lambda_i)$

Adding a layer



Coefficients of 10exp(-10t)

n=5

$a(10exp(-10t))$
1
$-6/4$
$9/14$
$-1/14$
$5/70$

*3 =

n=6

$a(10exp(-10t))$
3
$-27/6$
$81/42$
$-9/42$
$6/42$

n=5 → 6(i.e. add 15)

15/(15-10)=3

15/(15-6)=5/3

15/(15-3)=5/4

15/14

Table 1. n=5

	$\lambda's$	p_k	$a(10exp(-10t))$	$a(6exp(-6t))$	$a(3exp(-3t)$	$exp(-t)$
1	10	4/10	1			
2	10, 6	3/10	−6/4	10/4		
3	10, 6, 3	2/10	9/14	−35/14	40/14	
4	10, 60, 3, 1	1/10	−1/14	7/14	−20/14	28/14
	totals	1	5/70	21/70	30/70	14/70

n=6

Table 1b. n=6

k	$\lambda's$	p_k	$a(15exp(-15t))$	$a(10exp(-10t))$	$a(6exp(-6t))$	$a(3exp(-3t))$	$exp(-t)$
1	15	5/15	1				
2	15, 10	4/15	−2	3			
3	15, 10, 6	3/15	8/6	−27/6	25/6		
4	15, 10, 6, 3	2/15	−14/42	81/42	−175/42	150/42	
5	15, 10, 6, 3, 1	1/15	1/42	−9/42	35/42	−75/42	90/42
	totals	1	1/42	6/42	14/42	15/42	6/42

From sum row=1

Prob for Exponential with $\lambda = {}_kC_2$ for tree of depth n, *f(n,k)*

$$f(n,k) = \Sigma_{j=1}^{k-1} \frac{j}{{}_nC_2} \frac{\Pi_{u=j+1,u\neq k}^{n} {}_uC_2}{\Pi_{u=j+1,u\neq k}^{n} \{{}_uC_2 - {}_kC_2\}}$$

$$= \Pi_{u=k+1}^{n} \frac{{}_uC_2}{\{{}_uC_2 - {}_kC_2\}} \frac{{}_kC_2}{{}_nC_2} * \Sigma_{j=1}^{k-1} \frac{j}{{}_kC_2} \frac{\Pi_{u=j+1,u\neq k}^{k-1} {}_uC_2}{\Pi_{u=j+1,u\neq k}^{k-1} \{{}_uC_2 - {}_kC_2\}}$$

$$= \Pi_{u=k+1}^{n} \frac{{}_uC_2}{\{{}_uC_2 - {}_kC_2\}} \frac{{}_kC_2}{{}_nC_2} * f(k,k)$$

Prob for exp($_kC_2$) for n

$$f(n,k) = \Sigma_{j=1}^{k-1} \frac{j}{{}_nC_2} \boxed{\frac{\Pi_{u=j+1,u\neq k}^{n} {}_uC_2}{\Pi_{u=j+1,u\neq k}^{n} \{{}_uC_2 - {}_kC_2\}}}$$

$$= \Pi_{u=k+1}^{n} \frac{{}_uC_2}{\{{}_uC_2 - {}_kC_2\}} \frac{{}_kC_2}{{}_nC_2} * \Sigma_{j=1}^{k-1} \frac{j}{{}_kC_2} \frac{\Pi_{u=j+1,u\neq k}^{k-1} {}_uC_2}{\Pi_{u=j+1,u\neq k}^{k-1} \{{}_uC_2 - {}_kC_2\}}$$

$$= \Pi_{u=k+1}^{n} \frac{{}_uC_2}{\{{}_uC_2 - {}_kC_2\}} \frac{{}_kC_2}{{}_nC_2} * f(k,k)$$

Prob for exp($_kC_2$) for n

$$f(n,k) = \Sigma_{j=1}^{k-1} \frac{j}{{}_nC_2} \frac{\Pi_{u=j+1,u\neq k}^{n} {}_uC_2}{\Pi_{u=j+1,u\neq k}^{n} \{{}_uC_2 - {}_kC_2\}}$$

$$= \Pi_{u=k+1}^{n} \frac{{}_uC_2}{\{{}_uC_2 - {}_kC_2\}} \frac{{}_kC_2}{{}_nC_2} * \Sigma_{j=1}^{k-1} \frac{j}{{}_kC_2} \frac{\Pi_{u=j+1,u\neq k}^{k-1} {}_uC_2}{\Pi_{u=j+1,u\neq k}^{k-1} \{{}_uC_2 - {}_kC_2\}}$$

$$= \Pi_{u=k+1}^{n} \frac{{}_uC_2}{\{{}_uC_2 - {}_kC_2\}} \frac{{}_kC_2}{{}_nC_2} * f(k,k)$$

>0

f(k,k)=coefficient of exp(-$_kC_2$) for k

$$f(k,k) = \Sigma_{j=1}^{k-1} \frac{j}{{}_kC_2} \frac{\Pi_{u=j+1,u\neq k}^{k-1} {}_uC_2}{\Pi_{u=j+1,u\neq k}^{k-1} \{{}_uC_2 - {}_kC_2\}}$$

$$= \frac{1}{{}_kC_2} \frac{\Sigma_{j=1}^{k-1} j \Pi_{u=j+1,u\neq k}^{k-1} {}_uC_2 \Pi_{u=2}^{j} \{{}_uC_2 - {}_kC_2\}}{\Pi_{u=2,u\neq k}^{k-1} \{{}_uC_2 - {}_kC_2\}}$$

Denominator of *f(k,k)*

$$\Pi_{u=2}^{k-1}\{{}_kC_2 -_u C_2\} = \Pi_{u=1}^{k-2}\{{}_kC_2 -_{k-u} C_2\}$$

$$= \Pi_1^{k-2} u(2k-i-1)/2 = \frac{(k-2)!(2k-2)!}{k!2^{k-2}}$$

$$= \frac{[2(k-1)]!}{k!(k-1)!}\frac{(k-1)!(k-2)!}{2^{k-2}}$$

Denominator of *f(k,k)*

$$= \frac{[2(k-1)]!}{k!(k-1)!} \frac{(k-1)!(k-2)!}{2^{k-2}}$$

$$= c_n * \Delta_k$$

Where c_n is Catalan number and Δk products of triangular numbers

Numerator of *f(k,k)*

$$\Sigma_{j=1}^{k-1} {}_{j}\Pi_{u=j+1,u\neq k}^{k-1} {}_{u}C_2 \Pi_{u=2}^{j} \{ {}_{u}C_2 - {}_{k}C_2 \}$$

$$= \Delta_k$$

f(k,k)

- Thus f(k,k)=1/c_n.

- f(n,k) > 0 and we have convex combinations of exponentials.

References

References

Caliebe A, Neininger R, Krawczak M & Rösler U (2007) On the length distribution of external branches in coalescence trees: genetic diversity within species. *Theoret Pop Biol*, 72, 245-252.

Cannings C (1974) The latent roots of certain Markov chains arising in genetics: A new approach, I haploid models. *Adv Appl Prob* 6, 260-290.

Fu Y-X & Li W-H (1993) Statistical tests of neutrality of mutations. *Genetics* 133, 693-709.

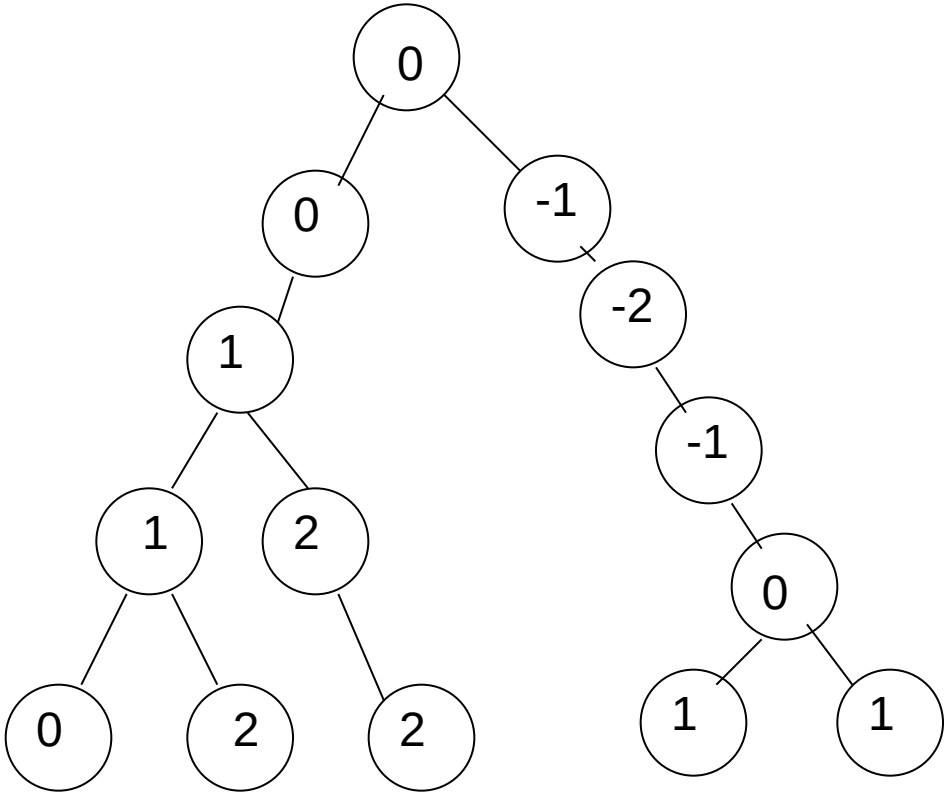
Kingman JFC (1982) The coalescent. *Stoch Proc and Appl*, 13, 2350248.

Nordborg M (2007) Coalescent theory. *Handbook of Statistical Genetics*, Wiley, Chichester. pgs 843-877. Editors DJ Balding, M Bishop, C Cannings.

Stepwise Mutation

- Suppose that we have variants of some gene which can be mapped to the integers (they may be the number of copies of some sequence [acct] [acct]…. [acct]). Suppose further that the copying from parent to offspring is subject to error so that each offspring produced by a parent with copy-number x will be (x-1), x or (x+1) with probabilities μ/2, (1- μ), μ/2.

Example



Stepwise Mutation

- How can we keep track of the distribution of values in a population of n undergoing reproduction as per the Wright-Fisher model?
- Suppose we start at time 0 with some value 0 (arbitrary). Now there will be n reproductions along the line from the founder to an individual in generation n. Thus the value will be simply $Z=\Sigma_{i=1,n}X_i$ where X_i is the step at i. Thus E(Z)=0, Var(Z)=nμ so that Var->inf as n->inf.

Stepwise Mutation

- Now consider the values normalised by some specified individual i.e. Z_i-Z_n. Now we need to know how many mutational events there have been between two individuals in generation n. Since there are N individuals in generation n then the chance that they share a parent in the previous generation is 1/N, and (1-1/N) that they have distinct parents.

Stepwise Mutation

- Thus the probability that they have a MRCA (most recent common ancestor) k generations ago is $(1-1/N)^{(k-1)}*1/N$ for k=1,2,…..,(n-1), and $(1-1/N)^{n-1}$ for k=n.
- The probability generating function for the random variable in question , where

$$x = (N-1)/N$$

$$g(s) = \frac{s(1-(xs)^{n-1})}{N(1-xs)} + s^n x^{n-1}$$

Stepwise Mutation

- Given k generations until join there will be 2(k-1) mutation “opportunities” so the pgf for this variable is $g(s^2)$.
- The pgf for the mutation step is

$$f(s) = (1-\mu) + s\mu + s^{-1}\mu$$

so we then examine g(f(s)) for the separation of the two individuals.

Moments of *K* with pgf *h(s)*

$$E(K_{[r]}) = [d^r h(s) / ds^r]_{s=1}$$

$$d^r g(f(s))/ds^r$$

- Here

$$h(s)=g(f(s))$$

$d^r g(f(s))/ds^r$

We write $g_i = \frac{d^i g}{ds^i}$ and similarly for f_i and h_i.

Now $h = g(f)$ so $h_1 = g_1 f_1$,

$$h_2 = g_2(f_1)^2 + g_1 f_2,$$

$$h_3 = g_3(f_1)^3 + 3g_2 f_1 f_2 + g_1 f_3,$$

$$h_4 = g_4(f_1)^4 + 6g_3(f_1)^2 f_2 + 3g_2(f_2)^2 + 4g_2 f_1 f_3 + g_1 f_4$$

$d^r g(f(s))/ds^r$

- Now each time we differentiate a g_i we get $g_{i+1}f_1$ and each time we differentiate an f_i we get an f_{i+1}. So each term is of the form

$$g_r f_1^{x_1} f_2^{x_2} \dots f_t^{x_t} \text{ where } \Sigma_{i=1}^{t} x_i = r$$

$d^r g(f(s))/ds^r$

We write $g_i = \frac{d^i g}{ds^i}$ and similarly for f_i and h_i.

Now $h = g(f)$ so $h_1 = g_1 f_1$, **1**

$h_2 = g_2(f_1)^2 + g_1 f_2$, **1,1 or 2**

$h_3 = g_3(f_1)^3 + 3g_2 f_1 f_2 + g_1 f_3$, **1,1,1 or 2,1 or 3**

$h_4 = g_4(f_1)^4 + 6g_3(f_1)^2 f_2 + 3g_2(f_2)^2 + 4g_2 f_1 f_3 + g_1 f_4$

1,1,1,1 or 2,1,1 or 2,2 or 3,1 or 4

$$d^r g(f(s))/ds^r$$

- We can see that the process of differentiating n times just corresponds to placing n balls in boxes. For example with n=5 the term 2,2,1 corresponds to $f_1(f_2)^2$ and then we add the

g_3 to give $g_3 f_1(f_2)^2$

- The number of terms correspond to labelled balls in boxes; Bell numbers/polynomials.

Bell numbers

- Suppose we want the coefficient for $g_3 f_1 (f_2)^2$

  then we need to know how many ways we can put 5 labelled balls into 3 unlabelled boxes with 1 in one box and 2 in each of the others; just 5 ways to choose the singleton and then ${}_4C_2/2$ to choose the two dividing by two since we do not distinguish the boxes, i.e. 15.

Stepwise Mutation

- Now since $f(s) = (1-\mu) + s\mu + s^{-1}\mu$

$$f(1) = 1,\ f_1(s) = \mu - \mu / s^2,\ f_1(1) = 0$$

Stepwise Mutation

- Thus the terms with f_1 will be eliminated from the expressions for the moments of K.
- The equivalent in terms of the balls in boxes is to not allow any box to have a single ball.

$d^r g(f(s))/ds^r$

We write $g_i = \frac{d^i g}{ds^i}$ and similarly for f_i and h_i.

Now $h = g(f)$ so $h_1 = g_1 f_1$,

$h_2 = g_2 (f_1)^2 + g_1 f_2$,

$h_3 = g_3 (f_1)^3 + 3 g_2 f_1 f_2 + g_1 f_3$,

$h_4 = g_4 (f_1)^4 + 6 g_3 (f_1)^2 f_2 + 3 g_2 (f_2)^2 + 4 g_2 f_1 f_3 + g_1 f_4$

All f_i=0 for i odd

$$d^r g(f(s))/ds^r$$

We can only include even *f's.* Thus we get

$$h_2 = g_1 f_2$$

$$h_4 = 3g_2 (f_2)^2 + g_1 f_4$$

$$h_6 = 15 g_3 (f_2)^3 + 15 g_2 f_2 f_4 + g_1 f_6$$

$$h_8 = 105 g_4 (f_2)^4 + 210 g_3 (f_2)^2 f_4 + 35 g_2 (f_4)^2 + 28 g_2 f_2 f_6 + g_1 f_8$$

Stepwise Mutation

- From the above theory we can easily obtain the moments of the random variable in question.
- We know the mean=0 and can prove the variance $=2\mu N(1-(1-1/N)^{n})$.

Higher Order Variables

- We might look at the time back from some number greater than two to their common ancestor.

Thoughts

- By reversing time we switch attention from a complex process to a relatively simple one on a tree.
- Different questions seem to throw up distinct tree topologies in an interesting way.