

Coalescence Theory and Population Genetics

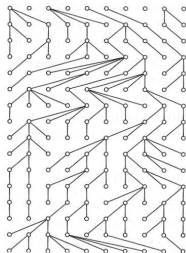
Ola Hössjer

May, 2011

Department of Mathematics, Stockholm University

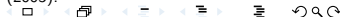
Wright Fisher Model

- Non-overlapping generations.
- N “individuals” in all generations (constant population size).
- Each individual chooses randomly parent from previous generation.



16 generations (time proceeds downwards), $N = 10$

This and some other pictures from Hein et al. (2005).

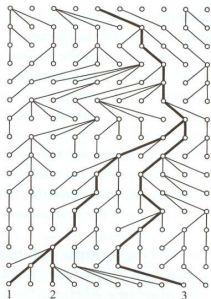


WF Backwards in Time

Given sample of n individuals, let

$$\begin{aligned} T_k &= \text{nr. of generations with } k \text{ ancestors} \\ &= \text{nr. of generations until some of } k \text{ lineages } \mathbf{coalesce} \end{aligned}$$

for $k = n, n - 1, \dots, 2$.



$$N = 10, n = 3, T_3 = 2, T_2 = 7.$$

Distribution of Coalescence Times

Assume N large, n fixed. Let

$$\begin{aligned} p_k &= P(\geq 2 \text{ of } k \text{ lineages coalesce in one generation}) \\ &= 1 - \prod_{i=1}^{k-1} (1 - i/N), \\ &= \binom{k}{2}/N + o(N^{-1}), \end{aligned}$$

Reproduction is independent between generations

\implies

Given $T_k > 0$, T_k has a **geometric distribution**

$$P(T_k = s) = (1 - p_k)^{s-1} p_k, \quad s = 1, 2, \dots$$

Coalescence in Continuous Time

Count time in units of N generations, with $\tau_k = T_k/N$. As $N \rightarrow \infty$,

$$\begin{aligned}P(\tau_k \leq t) &= P(T_k \leq Nt) \\&= 1 - (1 - p_k)^{[Nt]} \\&\sim 1 - \left(1 - \binom{k}{2}/N\right)^{[Nt]} \\&\rightarrow 1 - \exp\left(-\binom{k}{2}t\right)\end{aligned}$$

i.e. τ_k has exponential distribution with

$$E(\tau_k) = 1/\binom{k}{2}.$$

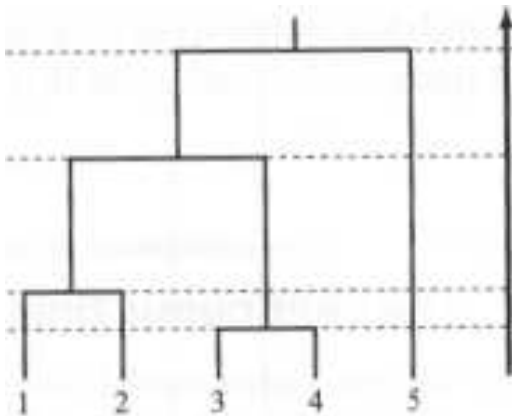
One says that

k ancestral lineages coalesce at **rate** $\binom{k}{2}$.

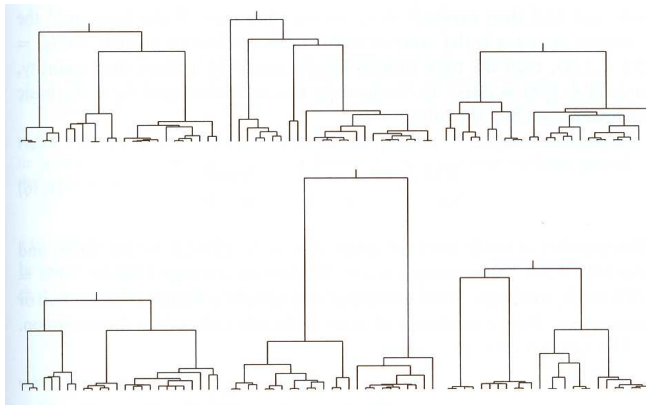
Continuous Time Coalescent

Coalescent (tree) specifies genealogy of the sample with **time on vertical axis**, so that all coalescent events are shown.

Coalescent tree with $n = 5$:



Six Simulated Cont Time Coalescents, $n = 25$



Coalescence rate and individual variability in reproductivity

Let

ν^l = number of children of parent $l = 1, \dots, N$

of a fixed generation. Constant population size

$$\sum_{l=1}^N \nu^l = N$$

Assume

$$\begin{aligned} E(\nu^l) &= 1, \\ \text{Var}(\nu^l) &= \lambda (\approx 1 \text{ for WF}). \end{aligned}$$

Probability

$$p_k = \lambda \binom{k}{2} / N + o(N^{-1})$$

that at least two of k lineages coalesce in one generation.

Continuous Time

Let $N \rightarrow \infty$, keep λ fixed.

Count time in units of

$$N_e = N/\lambda$$

generations. Gives same exponential distribution

$$\begin{aligned} P(\tau_k \leq t) &= P(T_k \leq Nt/\lambda) \\ &\sim 1 - (1 - \lambda \binom{k}{2} / N)^{[Nt/\lambda]} \\ &\rightarrow 1 - \exp(-\binom{k}{2} t). \end{aligned}$$

for τ_k as before.



“Coalescence time” running faster by factor λ .

Let

$$\begin{aligned} N_e &= N/\lambda \\ &= \text{effective population size} \\ &= \text{inverse of overall speed of coalescence} \\ &\quad \text{on original time scale (generations)} \\ &= \text{rate of loss of genetic variability} \end{aligned}$$

Large enough N_e required to avoid population extinction:

$$N_e \leq 50 \implies \text{inbreeding and genetic defects} \\ \text{within a few generations}$$

Diploid Wright-Fisher Model

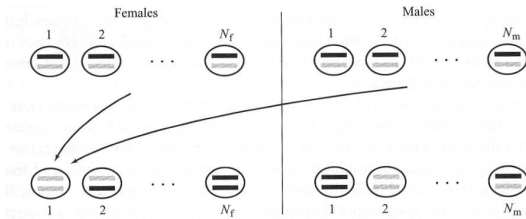
Constant populations size with

- N_m = number of males, N_f = number of females
- $N = 2(N_f + N_m)$ alleles (“individuals”) = twice nr. of individuals
- Allele = small stretch of DNA.
- $c = N_m / (N_m + N_f)$ proportion of males

Individuals (male or female) pick randomly

- paternal allele from one of $2N_m$ alleles,
- maternal allele from one of $2N_f$ alleles

of previous generation.



Coalescence in One Generation

The probability that 2 alleles coalesce in one generation is

$$\begin{aligned} p_2 &= \frac{0.5N(0.5N-1)}{N(N-1)} \left(\frac{1}{2N_m} + \frac{1}{2N_f} \right) \\ &= \lambda/N + o(N^{-1}), \end{aligned}$$

where

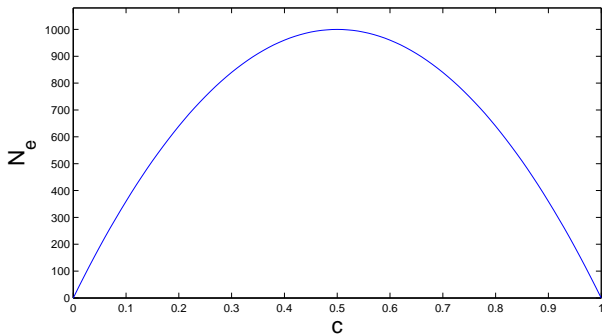
$$\begin{aligned} \lambda &= 1/(4c(1-c)) \\ N_e &= N/\lambda = 4c(1-c)N. \end{aligned}$$

More generally, probability

$$p_k = \lambda \binom{k}{2} / N + o(N^{-1})$$

that at least two of k lineages coalesce in one generation.

Effective Population Size, Diploid Population



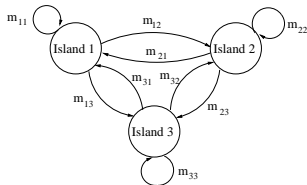
N_e versus c when $N = 1000$.

Find N_e for Populations Divided into Subpopulations

1. Geographical structure.

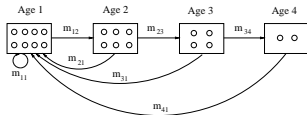
Subpopulation = geographical site (island, ...).

m_{ij} migration rate between subpopulations i and j .

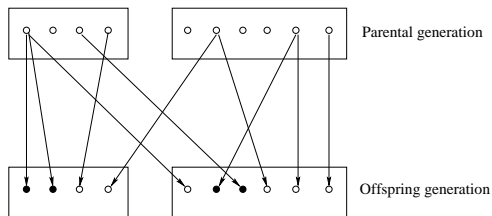


2. Age structure

Subpopulation = age class.



Two subpopulations and two generations



$$N = 10 = 4 + 6,$$

$$n = 4 \text{ (filled circles of offspring generation)}$$

$$T_4 = 1 \text{ (first two individuals of subpop 1 have same parent)}$$

Varying Population Size

One-sex Wright-Fisher model with

- N “individuals” at present time.
- Count time **backwards** in units of N generations
- $N(t)$ “individuals” at time t ($= [Nt]$ gener. back in time).
- Let $N \rightarrow \infty$, keep $\lambda(t) = N/N(t)$ fixed.
- **Time varying probability**

$$\begin{aligned} p_k(t) &= \binom{k}{2} / N(t) + o(N^{-1}) \\ &= \lambda(t) \binom{k}{2} / N + o(N^{-1}). \end{aligned}$$

that at least two of k lineages coalesce in **one** generation.

- On continuous time scale

k lineages coalesce at **time varying rate** $\lambda(t)$.

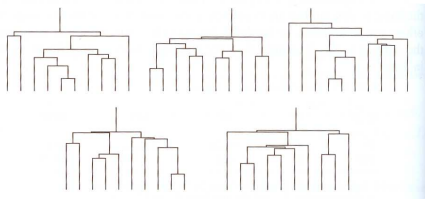
Exponential Population Growth

Let

$$\begin{aligned}N(t) &= N \exp(-\beta t), \\ \lambda(t) &= \exp(\beta t)\end{aligned}$$

where

$$\begin{aligned}\beta &= Nb = \text{scaled growth rate,} \\ b &= \text{per generation growth rate.}\end{aligned}$$



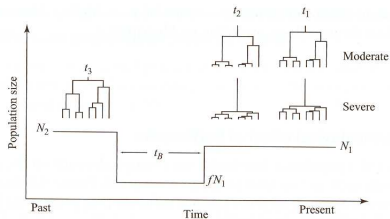
Five simulations, $n = 10$ and $\beta = 1000$.

Longer terminal branches than ordinary coalescent.

Star topology as $\beta \rightarrow \infty$ (all n lineages coalesce at once).

Population Bottleneck

$$N(t) = \begin{cases} N = N_1, & 0 \leq t \leq t_2, \\ fN_1, & t_2 \leq t \leq t_2 + t_B, \\ N_2, & t_2 + t_B \leq t. \end{cases}$$



Coalescence tree starting at time

- $t_1 = t_2 - \varepsilon$: Some time “before” bottleneck.
- t_2 : Just “before” bottleneck.
- $t_3 = t_2 + t_B$: Just “after” bottleneck.

Bertoin J. (2006). *Random fragmentation and coagulation processes*. Cambridge studies in advance mathematics.

Durrett, R. (2008). *Probability models for DNA sequence evolution*. 2nd edition. Probability and its applications, Springer, New York.

Ewens, W.J. (2004). *Mathematical models of population genetics. I. Theoretical introduction*. 2nd edition. Springer, New York.

Hein, J., Schierup, M.H. and Wiuf, C. (2005). *Gene genealogies, variation and evolution*. Oxford University Press, Oxford.

Hössjer, O. (2011). Coalescence theory for a general class of structured populations with fast migration. Research report 2011:2, Mathematical Statistics, Stockholm University.