UPPSALA
UNIVERSITET

HOW ARE WE RELATED TO EACH OTHER?

SIMPLE MODELS OF OUR GENETIC ANCESTRY

DOCENTSHIP LECTURE BY RAAZESH SAINUDIIN
DEPARTMENT OF MATHEMATICS, UPPSALA UNIVERSITY
2022-10-07

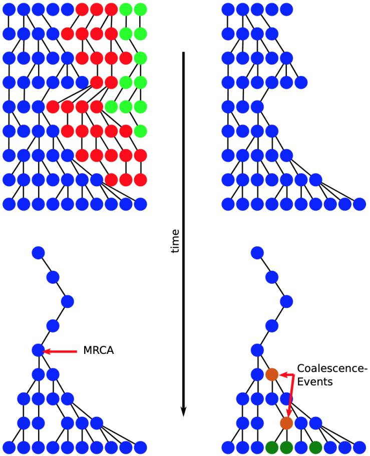
UPPSALA
UNIVERSITET

OUTLINE

- Introduce the Wright-Fisher model.
- Genealogies in the Wright-Fisher model.
- How to obtain the coalescent from the Wright-Fisher model.
- Properties of coalescent models

OUTLINE

- Introduce the Wright-Fisher model.
- Genealogies in the Wright-Fisher model.
- How to obtain the coalescent from the Wright-Fisher model.
- Properties of coalescent models

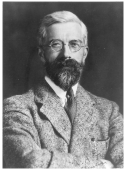


Wright-Fisher and Coalescent Models

THE WRIGHT-FISHER MODEL: INTRODUCTION

- There are *N* individuals in each generation
- Sample current generation with replacement to produce next generation
- Repeat the above step every generation

Sewall Wright
(1889-1988)

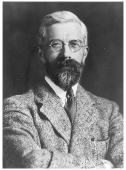


Sir Ronald Fisher
(1890-1962)

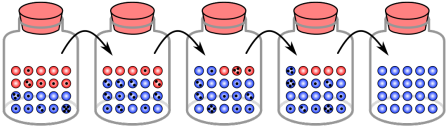
THE WRIGHT-FISHER MODEL: INTRODUCTION

- There are *N* individuals in each generation
- Sample current generation with replacement to produce next generation
- Repeat the above step every generation

Sewall Wright
(1889-1988)



Sir Ronald Fisher
(1890-1962)

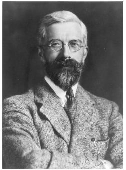


“Think of choosing each of the $N = 20$ balls in one jar (current generation) with the same probability $1/20$, independently 20 times for copying (reproduction) it to the next jar (next generation)”

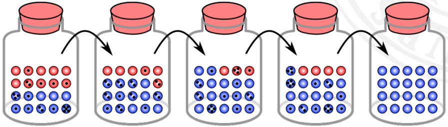
THE WRIGHT-FISHER MODEL: INTRODUCTION

- There are N individuals in each generation
- Sample current generation with replacement to produce next generation
- Repeat the above step every generation

Sewall Wright
(1889-1988)



Sir Ronald Fisher
(1890-1962)



“Think of choosing each of the $N = 20$ balls in one jar (current generation) with the same probability $1/20$, independently 20 times for copying (reproduction) it to the next jar (next generation)”

Theorem (Offspring Distribution)

This reproduction mechanism of the W-F Model gives the number of offspring in the next generation for each individual in the current generation labelled by $1, 2, \ldots, N$ *as the* $\mathrm{Multinomial}(N, (1/N, 1/N, \ldots, 1/N))$ *distribution .*

Mathematical Preliminaries 0.1 – basic counting

Lemma (number of permutations of n elements)

The number of different ways of arranging n distinct elements into a sequence*, a.k.a., the permutations of n, is* $n! := n \cdot (n-1) \cdot (n-2) \cdots 2 \cdot 1$*.*

Mathematical Preliminaries 0.1 – basic counting

Lemma (number of permutations of *n* elements)

The number of different ways of arranging n distinct elements into a sequence*, a.k.a., the permutations of n, is* $n! := n \cdot (n-1) \cdot (n-2) \cdots 2 \cdot 1$*.*

For example, there are $3! = 3 \cdot 2 \cdot 1 = 6$ permutations of $\{1,2,3\}$. They are:

$$(1,2,3),(2,1,3),(3,1,2),(1,3,2),(2,3,1),(3,2,1).$$

Mathematical Preliminaries 0.1 – basic counting

Lemma (number of permutations of *n* elements)

The number of different ways of arranging n distinct elements into a sequence*, a.k.a., the permutations of n, is* $n! := n\cdot(n-1)\cdot(n-2)\cdots 2\cdot 1$*.*

For example, there are $3! = 3\cdot 2\cdot 1 = 6$ permutations of $\{1,2,3\}$. They are:

$$(1,2,3),(2,1,3),(3,1,2),(1,3,2),(2,3,1),(3,2,1).$$

Proof.

The proof follows from the the following construction:

- there are n possible choices for the first element in the sequence
- for each such choice there are $(n-1)$ elements left to choose from for the second element
- and so on, until there is exactly one element left for the n-th element in the sequence
- the total number of permutations follow from multiplying the number of possible choices in each of the n steps above giving $n\cdot(n-1)\cdot(n-2)\cdots 2\cdot 1$.
- Note that there is only one way of arranging $\{\}$, an empty set without any element, so $0! = 1$.

□

Mathematical Preliminaries 0.2 – basic counting

Lemma (number of subsets of size k from a set with n elements)

The number of the k-element combinations, a.k.a., subsets of k elements from a set with n elements, is

$$\binom{n}{k} := \frac{n!}{k!(n-k)!} = \frac{n \cdot (n-1) \cdot (n-2) \cdots (n-k+1)}{k \cdot (k-1) \cdots 2 \cdot 1}$$

Mathematical Preliminaries 0.2 – basic counting

Lemma (number of subsets of size *k* from a set with *n* elements)

The number of the k-element combinations, a.k.a., subsets of k elements from a set with n elements, is

$$\binom{n}{k} := \frac{n!}{k!(n-k)!} = \frac{n\cdot(n-1)\cdot(n-2)\cdots(n-k+1)}{k\cdot(k-1)\cdots 2\cdot 1}$$

For example, there are $\frac{5\cdot4\cdot3}{3\cdot2\cdot1} = \frac{60}{6} = 10$ 3-element combinations of $\{1,2,3,4,5\}$. They are:

$$\{1,2,3\},\{1,2,4\},\{1,2,5\},\{1,3,4\},\{1,3,5\},\{1,4,5\},\{2,3,4\},\{2,3,5\},\{2,4,5\},\{3,4,5\}.$$

Mathematical Preliminaries 0.2 – basic counting

Lemma (number of subsets of size *k* from a set with *n* elements)

The number of the k-element combinations, a.k.a., subsets of k elements from a set with n elements, is

$$\binom{n}{k} := \frac{n!}{k!(n-k)!} = \frac{n\cdot(n-1)\cdot(n-2)\cdots(n-k+1)}{k\cdot(k-1)\cdots 2\cdot 1}$$

For example, there are $\frac{5\cdot 4\cdot 3}{3\cdot 2\cdot 1} = \frac{60}{6} = 10$ 3-element combinations of $\{1,2,3,4,5\}$. They are:

$$\{1,2,3\},\{1,2,4\},\{1,2,5\},\{1,3,4\},\{1,3,5\},\{1,4,5\},\{2,3,4\},\{2,3,5\},\{2,4,5\},\{3,4,5\}.$$

Proof.

Follows by construction where,

- the numerator counts the number of ways to choose the first *k* elements from the set of *n* elements and arrange them in a row with regard to order
- the denominator disregards order as each *k*-element combination appears exactly *k*! times

□

THE WRIGHT-FISHER MODEL: MULTINOMIAL OFFSPRING

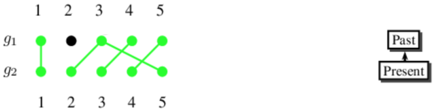
- Label individuals in (parent) generation g_1 by $1,2,3,4,5$ in the W-F Model with $N=5$

THE WRIGHT-FISHER MODEL: MULTINOMIAL OFFSPRING

- Label individuals in (parent) generation g_1 by $1, 2, 3, 4, 5$ in the W-F Model with $N = 5$
- Label individuals in (offspring) generation g_2 by $1, 2, 3, 4, 5$
- Let each individual in g_2 chooses a parent in g_1 independently with probability $\frac{1}{N} = \frac{1}{5}$

THE WRIGHT-FISHER MODEL: MULTINOMIAL OFFSPRING

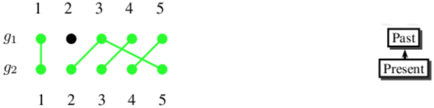
- Label individuals in (parent) generation g_1 by $1,2,3,4,5$ in the W-F Model with $N=5$
- Label individuals in (offspring) generation g_2 by $1,2,3,4,5$
- Let each individual in g_2 chooses a parent in g_1 independently with probability $\frac{1}{N}=\frac{1}{5}$
- Consider the offspring-parent relation $(1,2,3,4,5) \overset{\text{parent}}{\longrightarrow} (1,3,4,5,3)$



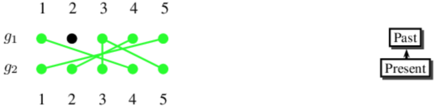
$$\Pr\left\{(1,2,3,4,5) \overset{\text{parent}}{\longrightarrow} (1,3,4,5,3)\right\} = \left(\frac{1}{5}\right)^5$$

THE WRIGHT-FISHER MODEL: MULTINOMIAL OFFSPRING

- Label individuals in (parent) generation g_1 by $1,2,3,4,5$ in the W-F Model with $N = 5$
- Label individuals in (offspring) generation g_2 by $1,2,3,4,5$
- Let each individual in g_2 chooses a parent in g_1 independently with probability $\frac{1}{N} = \frac{1}{5}$
- Consider the offspring-parent relation $(1,2,3,4,5) \overset{\text{parent}}{\longrightarrow} (1,3,4,5,3)$
- Consider another offspring-parent relation $(1,2,3,4,5) \overset{\text{parent}}{\longrightarrow} (5,4,3,1,3)$

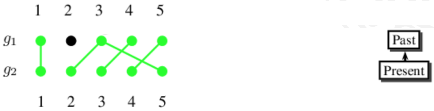


$$\Pr\left\{(1,2,3,4,5) \overset{\text{parent}}{\longrightarrow} (1,3,4,5,3)\right\} = \left(\frac{1}{5}\right)^5$$

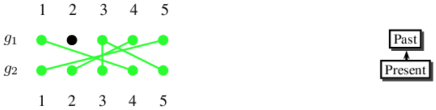


THE WRIGHT-FISHER MODEL: MULTINOMIAL OFFSPRING

- Label individuals in (parent) generation g_1 by $1,2,3,4,5$ in the W-F Model with $N=5$
- Label individuals in (offspring) generation g_2 by $1,2,3,4,5$
- Let each individual in g_2 chooses a parent in g_1 independently with probability $\frac{1}{N}=\frac{1}{5}$
- Consider the offspring-parent relation $(1,2,3,4,5) \overset{\text{parent}}{\longrightarrow} (1,3,4,5,3)$
- Consider another offspring-parent relation $(1,2,3,4,5) \overset{\text{parent}}{\longrightarrow} (5,4,3,1,3)$



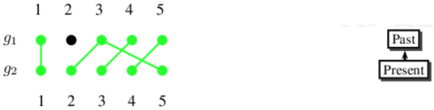
$$\Pr\left\{(1,2,3,4,5) \overset{\text{parent}}{\longrightarrow} (1,3,4,5,3)\right\} = \left(\frac{1}{5}\right)^5$$



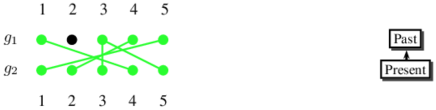
$$\Pr\left\{(1,2,3,4,5) \overset{\text{parent}}{\longrightarrow} (5,4,3,1,3)\right\} = \left(\frac{1}{5}\right)^5$$

THE WRIGHT-FISHER MODEL: MULTINOMIAL OFFSPRING

- Label individuals in (parent) generation g_1 by $1,2,3,4,5$ in the W-F Model with $N=5$
- Label individuals in (offspring) generation g_2 by $1,2,3,4,5$
- Let each individual in g_2 chooses a parent in g_1 independently with probability $\frac{1}{N}=\frac{1}{5}$
- Consider the offspring-parent relation $(1,2,3,4,5) \xrightarrow{\text{parent}} (1,3,4,5,3)$
- Consider another offspring-parent relation $(1,2,3,4,5) \xrightarrow{\text{parent}} (5,4,3,1,3)$
- Probability of any of the N^N many offspring-parent relations is $\left(\frac{1}{N}\right)^N$



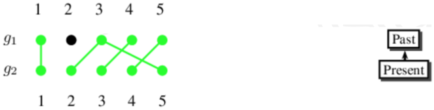
$$\Pr\left\{(1,2,3,4,5) \xrightarrow{\text{parent}} (1,3,4,5,3)\right\} = \left(\frac{1}{5}\right)^5$$



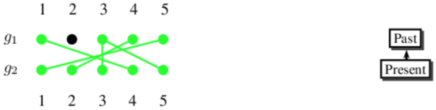
$$\Pr\left\{(1,2,3,4,5) \xrightarrow{\text{parent}} (5,4,3,1,3)\right\} = \left(\frac{1}{5}\right)^5$$

THE WRIGHT-FISHER MODEL: MULTINOMIAL OFFSPRING

- Label individuals in (parent) generation g_1 by $1,2,3,4,5$ in the W-F Model with $N=5$
- Label individuals in (offspring) generation g_2 by $1,2,3,4,5$
- Let each individual in g_2 chooses a parent in g_1 independently with probability $\frac{1}{N}=\frac{1}{5}$
- Consider the offspring-parent relation $(1,2,3,4,5)\overset{\text{parent}}{\longrightarrow}(1,3,4,5,3)$
- Consider another offspring-parent relation $(1,2,3,4,5)\overset{\text{parent}}{\longrightarrow}(5,4,3,1,3)$
- Probability of any of the N^N many offspring-parent relations is $\left(\frac{1}{N}\right)^N$
- Let X_i be offspring number of parent i. What is $\Pr(X_1=x_1,X_2=x_2,\dots X_N=x_N)$?



$$\Pr\left\{(1,2,3,4,5)\overset{\text{parent}}{\longrightarrow}(1,3,4,5,3)\right\}=\left(\frac{1}{5}\right)^5$$

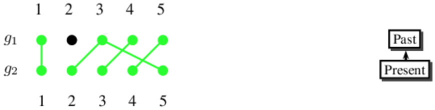


$$\Pr\left\{(1,2,3,4,5)\overset{\text{parent}}{\longrightarrow}(5,4,3,1,3)\right\}=\left(\frac{1}{5}\right)^5$$

$$\Pr(X_1=1,X_2=0,X_3=2,X_4=1,X_5=1)\geq 2\left(\frac{1}{5}\right)^5$$

But what is it exactly?

THE WRIGHT-FISHER MODEL: MULTINOMIAL OFFSPRING



$$\Pr\left\{(1,2,3,4,5) \xrightarrow{\text{parent}} (1,3,4,5,3)\right\} = \left(\frac{1}{5}\right)^5$$

$$\Pr\left\{(1,2,3,4,5) \xrightarrow{\text{parent}} (5,4,3,1,3)\right\} = \left(\frac{1}{5}\right)^5$$

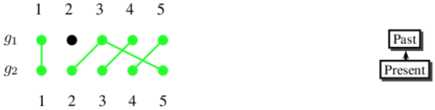
$$\Pr(X_1 = 1, X_2 = 0, X_3 = 2, X_4 = 1, X_5 = 1) = ?? \left(\frac{1}{5}\right)^5$$

Lemma (Multinomial Coefficient)

Let $\binom{N}{x_1, x_2, \ldots, x_N}$ *be* the number of ways parent i chooses x_i many offspring, $\forall i = 1, 2, \ldots, N$. *Then*

$$\binom{N}{x_1, x_2, \ldots, x_N} = \frac{N!}{x_1! x_2! \cdots x_N!}$$

THE WRIGHT-FISHER MODEL: MULTINOMIAL OFFSPRING



$$\Pr\left\{(1,2,3,4,5) \xrightarrow{\text{parent}} (1,3,4,5,3)\right\} = \left(\frac{1}{5}\right)^5$$

$$\Pr\left\{(1,2,3,4,5) \xrightarrow{\text{parent}} (5,4,3,1,3)\right\} = \left(\frac{1}{5}\right)^5$$

$$\Pr(X_1 = 1, X_2 = 0, X_3 = 2, X_4 = 1, X_5 = 1) = ??\left(\frac{1}{5}\right)^5$$

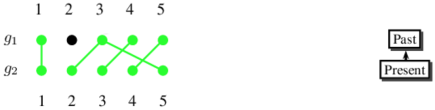
Lemma (Multinomial Coefficient)

Let $\binom{N}{x_1, x_2, \ldots, x_N}$ *be* the number of ways parent i chooses x_i many offspring, $\forall i = 1, 2, \ldots, N$. *Then*

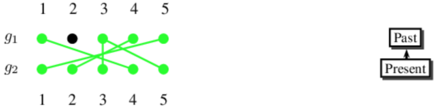
$$\binom{N}{x_1, x_2, \ldots, x_N} = \frac{N!}{x_1! x_2! \cdots x_N!}$$

- Before proving the Lemma let's apply it to our example

THE WRIGHT-FISHER MODEL: MULTINOMIAL OFFSPRING



$$\Pr\left\{(1,2,3,4,5) \xrightarrow{\text{parent}} (1,3,4,5,3)\right\} = \left(\frac{1}{5}\right)^5$$



$$\Pr\left\{(1,2,3,4,5) \xrightarrow{\text{parent}} (5,4,3,1,3)\right\} = \left(\frac{1}{5}\right)^5$$

$$\Pr(X_1 = 1, X_2 = 0, X_3 = 2, X_4 = 1, X_5 = 1) = ?? \left(\frac{1}{5}\right)^5$$

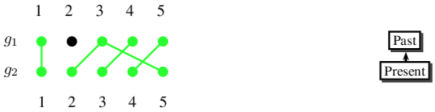
Lemma (Multinomial Coefficient)

Let $\binom{N}{x_1, x_2, \ldots, x_N}$ *be* the number of ways parent i chooses x_i many offspring, $\forall i = 1, 2, \ldots, N$. *Then*

$$\binom{N}{x_1, x_2, \ldots, x_N} = \frac{N!}{x_1! x_2! \cdots x_N!}$$

- Before proving the Lemma let's apply it to our example
- $\binom{5}{1,0,2,1,1} = \frac{5!}{1!0!2!1!1!} = 60$

THE WRIGHT-FISHER MODEL: MULTINOMIAL OFFSPRING



$$\Pr\left\{(1,2,3,4,5) \xrightarrow{\text{parent}} (1,3,4,5,3)\right\} = \left(\frac{1}{5}\right)^5$$

$$\Pr\left\{(1,2,3,4,5) \xrightarrow{\text{parent}} (5,4,3,1,3)\right\} = \left(\frac{1}{5}\right)^5$$

$$\Pr(X_1 = 1, X_2 = 0, X_3 = 2, X_4 = 1, X_5 = 1) = 60\left(\frac{1}{5}\right)^5$$

Lemma (Multinomial Coefficient)

Let $\binom{N}{x_1,x_2,\ldots,x_N}$ *be* the number of ways parent i chooses x_i many offspring, $\forall i = 1,2,\ldots,N$. *Then*

$$\binom{N}{x_1,x_2,\ldots,x_N} = \frac{N!}{x_1!x_2!\cdots x_N!}$$

- Before proving the Lemma let's apply it to our example
- $\binom{5}{1,0,2,1,1} = \frac{5!}{1!0!2!1!1!} = 60$
- If the lemma were true, then there are 60 possible distinct parent-offspring relations such that parent 1 has 1 offspring, parent 2 has none, parent 3 has 2 offspring, and parent 4 has 1 offspring and parent 5 has 1 offspring.

THE WRIGHT-FISHER MODEL: MULTINOMIAL OFFSPRING

Lemma (Multinomial Coefficient)

Let $\binom{N}{x_1, x_2, \ldots, x_N}$ *be* the number of ways parent i chooses x_i many offspring, $\forall i = 1, 2, \ldots, N$. *Then*

$$\binom{N}{x_1, x_2, \ldots, x_N} = \frac{N!}{x_1! x_2! \cdots x_N!}$$

THE WRIGHT-FISHER MODEL: MULTINOMIAL OFFSPRING

Lemma (Multinomial Coefficient)

Let $\binom{N}{x_1, x_2, \ldots, x_N}$ *be* the number of ways parent i chooses x_i many offspring, $\forall i = 1, 2, \ldots, N$. *Then*

$$\binom{N}{x_1, x_2, \ldots, x_N} = \frac{N!}{x_1! x_2! \cdots x_N!}$$

Proof.

The proof follows from the following construction:

$\square$

THE WRIGHT-FISHER MODEL: MULTINOMIAL OFFSPRING

Lemma (Multinomial Coefficient)

Let $\binom{N}{x_1, x_2, \dots, x_N}$ *be* the number of ways parent i chooses x_i many offspring, $\forall i = 1, 2, \dots, N$. *Then*

$$\binom{N}{x_1, x_2, \dots, x_N} = \frac{N!}{x_1! x_2! \cdots x_N!}$$

Proof.

The proof follows from the following construction:

- Choosing x_1 individuals of the total N (in generation g_2) to be labelled as offspring of individual 1 (in generation g_1). This can be done in $\binom{N}{x_1}$ ways.

$\square$

THE WRIGHT-FISHER MODEL: MULTINOMIAL OFFSPRING

Lemma (Multinomial Coefficient)

Let $\binom{N}{x_1,x_2,\dots,x_N}$ *be* the number of ways parent i chooses x_i many offspring, $\forall i = 1,2,\dots,N$. *Then*

$$\binom{N}{x_1,x_2,\dots,x_N} = \frac{N!}{x_1!x_2!\cdots x_N!}$$

Proof.

The proof follows from the following construction:

- Choosing x_1 individuals of the total N (in generation g_2) to be labelled as offspring of individual 1 (in generation g_1). This can be done in $\binom{N}{x_1}$ ways.
- From the remaining $N - x_1$ individuals, choose x_2 to be labelled as offspring of individual 2. This can be done $\binom{N-x_1}{x_2}$ ways.

□

UPPSALA
UNIVERSITET

THE WRIGHT-FISHER MODEL: MULTINOMIAL OFFSPRING

Lemma (Multinomial Coefficient)

Let $\binom{N}{x_1, x_2, \ldots, x_N}$ *be* the number of ways parent i chooses x_i many offspring, $\forall i = 1, 2, \ldots, N$. *Then*

$$\binom{N}{x_1, x_2, \ldots, x_N} = \frac{N!}{x_1!x_2!\cdots x_N!}$$

Proof.

The proof follows from the following construction:

- Choosing x_1 individuals of the total N (in generation g_2) to be labelled as offspring of individual 1 (in generation g_1). This can be done in $\binom{N}{x_1}$ ways.
- From the remaining $N - x_1$ individuals, choose x_2 to be labelled as offspring of individual 2. This can be done $\binom{N-x_1}{x_2}$ ways.
- From the remaining $N - x_1 - x_2$ individuals choose x_3 to be labelled as offspring of individual 3. Again, this can be done $\binom{N-x_1-x_2}{x_3}$ ways. [Proof continued...]

□

THE WRIGHT-FISHER MODEL: MULTINOMIAL OFFSPRING

Lemma (Multinomial Coefficient)

Let $\binom{N}{x_1, x_2, \ldots, x_N}$ *be* the number of ways parent i chooses x_i many offspring, $\forall i = 1, 2, \ldots, N$. *Then*

$$\binom{N}{x_1, x_2, \ldots, x_N} = \frac{N!}{x_1! x_2! \cdots x_N!}$$

Proof.

The proof continues from the above construction:

$\square$

THE WRIGHT-FISHER MODEL: MULTINOMIAL OFFSPRING

Lemma (Multinomial Coefficient)

Let $\binom{N}{x_1, x_2, \ldots, x_N}$ *be* the number of ways parent i chooses x_i many offspring, $\forall i = 1, 2, \ldots, N$. *Then*

$$\binom{N}{x_1, x_2, \ldots, x_N} = \frac{N!}{x_1! x_2! \cdots x_N!}$$

Proof.

The proof continues from the above construction:

- Multiplying the number of choices at each step results in:

$$\binom{N}{x_1}\binom{N - x_1}{x_2}\binom{N - x_1 - x_2}{x_3} \cdots = \frac{N!}{(N - x_1)!} \frac{(N - x_1)!}{(N - x_1 - X_2)! x_2!} \frac{(N - x_1 - X_2)!}{(N - x_1 - X_2 - x_3)! x_3!} \cdots$$

$\square$

THE WRIGHT-FISHER MODEL: MULTINOMIAL OFFSPRING

Lemma (Multinomial Coefficient)

Let $\binom{N}{x_1, x_2, \ldots, x_N}$ *be* the number of ways parent i chooses x_i many offspring, $\forall i = 1, 2, \ldots, N$. *Then*

$$\binom{N}{x_1, x_2, \ldots, x_N} = \frac{N!}{x_1! x_2! \cdots x_N!}$$

Proof.

The proof continues from the above construction:

- Multiplying the number of choices at each step results in:

$$\binom{N}{x_1}\binom{N - x_1}{x_2}\binom{N - x_1 - x_2}{x_3} \cdots = \frac{N!}{(N - x_1)!} \frac{(N - x_1)!}{(N - x_1 - X_2)! x_2!} \frac{(N - x_1 - X_2)!}{(N - x_1 - X_2 - x_3)! x_3!} \cdots$$

- Cancellation results in the desired formula:

$$\frac{N!}{x_1! x_2! \cdots x_N!}.$$

□

THE WRIGHT-FISHER MODEL: MULTINOMIAL OFFSPRING

Combining the last two observations and lemmas, we have proved the following theorem.

Theorem (Offspring Distribution of W-F Model is Multinomial)

Let X_i be the number of offspring of parent i, $\forall i = 1, 2, \ldots, N$. Then,

$$\Pr(X_1 = x_1, X_2 = x_2, \ldots X_N = x_N) = \binom{N}{x_1, x_2, \ldots, x_N} \left(\frac{1}{N}\right)^N$$

where,

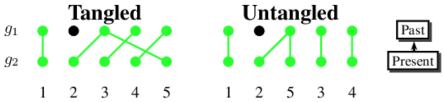
$$\binom{N}{x_1, x_2, \ldots, x_N} = \frac{N!}{x_1!x_2!\cdots x_N!}$$

THE WRIGHT-FISHER GENEALOGY

- The W-F Genealogy tracks the parent-offspring relations in the W-F Model

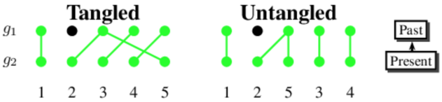
THE WRIGHT-FISHER GENEALOGY

- The W-F Genealogy tracks the parent-offspring relations in the W-F Model
- Let individuals in each generation be vertices and allow edges from parent to offspring vertices (the same graph is drawn in its “tangled” and “untangled” forms when $N = 5$)



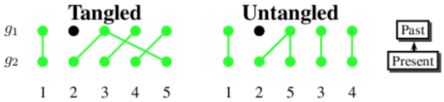
THE WRIGHT-FISHER GENEALOGY

- The W-F Genealogy tracks the parent-offspring relations in the W-F Model
- Let individuals in each generation be vertices and allow edges from parent to offspring vertices (the same graph is drawn in its “tangled” and “untangled” forms when $N = 5$)
- The graph with such vertices and edges is called the Wright-Fisher Genealogy



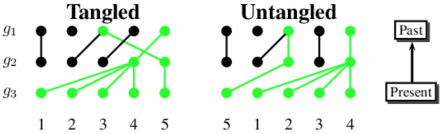
THE WRIGHT-FISHER GENEALOGY

- The W-F Genealogy tracks the parent-offspring relations in the W-F Model
- Let individuals in each generation be vertices and allow edges from parent to offspring vertices (the same graph is drawn in its "tangled" and "untangled" forms when $N = 5$)
- The graph with such vertices and edges is called the Wright-Fisher Genealogy
- The genealogy of a sample from the present population, i.e., sample genealogy (shown in green) is of interest



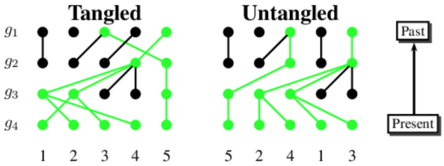
THE WRIGHT-FISHER GENEALOGY

- The W-F Genealogy tracks the parent-offspring relations in the W-F Model
- Let individuals in each generation be vertices and allow edges from parent to offspring vertices (the same graph is drawn in its “tangled” and “untangled” forms when $N = 5$)
- The graph with such vertices and edges is called the Wright-Fisher Genealogy
- The genealogy of a sample from the present population, i.e., sample genealogy (shown in green) is of interest



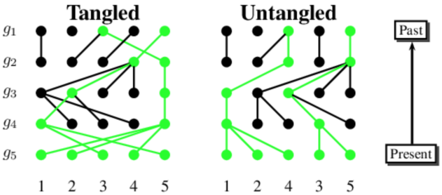
THE WRIGHT-FISHER GENEALOGY

- The W-F Genealogy tracks the parent-offspring relations in the W-F Model
- Let individuals in each generation be vertices and allow edges from parent to offspring vertices (the same graph is drawn in its “tangled” and “untangled” forms when $N = 5$)
- The graph with such vertices and edges is called the Wright-Fisher Genealogy
- The genealogy of a sample from the present population, i.e., sample genealogy (shown in green) is of interest



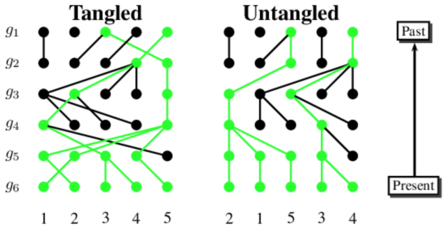
THE WRIGHT-FISHER GENEALOGY

- The W-F Genealogy tracks the parent-offspring relations in the W-F Model
- Let individuals in each generation be vertices and allow edges from parent to offspring vertices (the same graph is drawn in its “tangled” and “untangled” forms when $N = 5$)
- The graph with such vertices and edges is called the Wright-Fisher Genealogy
- The genealogy of a sample from the present population, i.e., sample genealogy (shown in green) is of interest



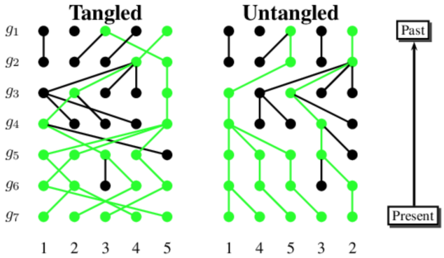
THE WRIGHT-FISHER GENEALOGY

- The W-F Genealogy tracks the parent-offspring relations in the W-F Model
- Let individuals in each generation be vertices and allow edges from parent to offspring vertices (the same graph is drawn in its "tangled" and "untangled" forms when $N = 5$)
- The graph with such vertices and edges is called the Wright-Fisher Genealogy
- The genealogy of a sample from the present population, i.e., sample genealogy (shown in green) is of interest



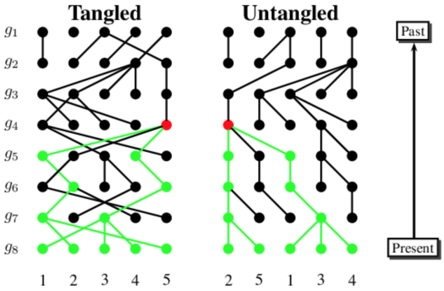
THE WRIGHT-FISHER GENEALOGY

- The W-F Genealogy tracks the parent-offspring relations in the W-F Model
- Let individuals in each generation be vertices and allow edges from parent to offspring vertices (the same graph is drawn in its “tangled” and “untangled” forms when $N = 5$)
- The graph with such vertices and edges is called the Wright-Fisher Genealogy
- The genealogy of a sample from the present population, i.e., sample genealogy (shown in green) is of interest



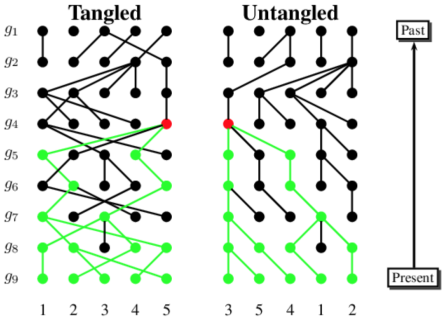
THE WRIGHT-FISHER GENEALOGY

- The W-F Genealogy tracks the parent-offspring relations in the W-F Model
- Let individuals in each generation be vertices and allow edges from parent to offspring vertices (the same graph is drawn in its “tangled” and “untangled” forms when $N = 5$)
- The graph with such vertices and edges is called the Wright-Fisher Genealogy
- The genealogy of a sample from the present population, i.e., sample genealogy (shown in green) is of interest
- The Most Recent Common Ancestor (MRCA) of the sample from the present (shown in red)



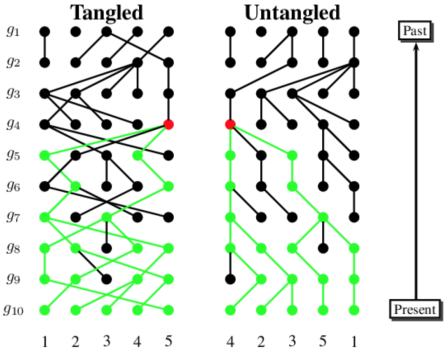
THE WRIGHT-FISHER GENEALOGY

- The W-F Genealogy tracks the parent-offspring relations in the W-F Model
- Let individuals in each generation be vertices and allow edges from parent to offspring vertices (the same graph is drawn in its “tangled” and “untangled” forms when $N = 5$)
- The graph with such vertices and edges is called the Wright-Fisher Genealogy
- The genealogy of a sample from the present population, i.e., sample genealogy (shown in green) is of interest
- The Most Recent Common Ancestor (MRCA) of the sample from the present (shown in red)

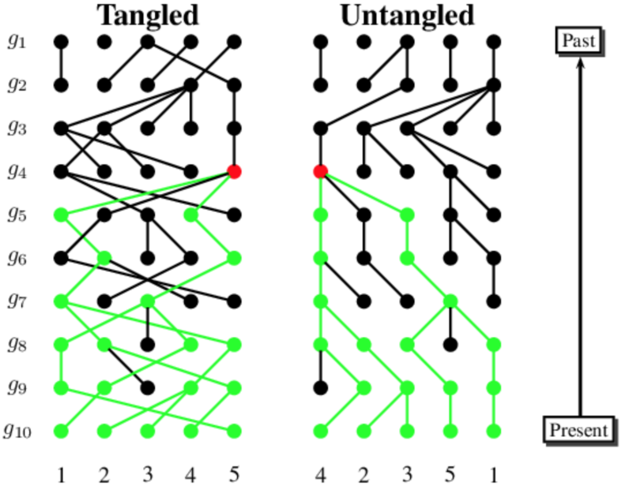


THE WRIGHT-FISHER GENEALOGY

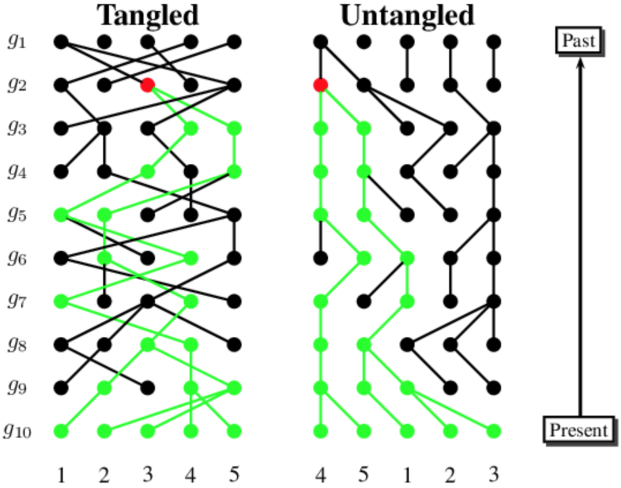
- The W-F Genealogy tracks the parent-offspring relations in the W-F Model
- Let individuals in each generation be vertices and allow edges from parent to offspring vertices (the same graph is drawn in its “tangled” and “untangled” forms when $N = 5$)
- The graph with such vertices and edges is called the Wright-Fisher Genealogy
- The genealogy of a sample from the present population, i.e., sample genealogy (shown in green) is of interest
- The Most Recent Common Ancestor (MRCA) of the sample from the present (shown in red)
- We want to study properties of the sample genealogy



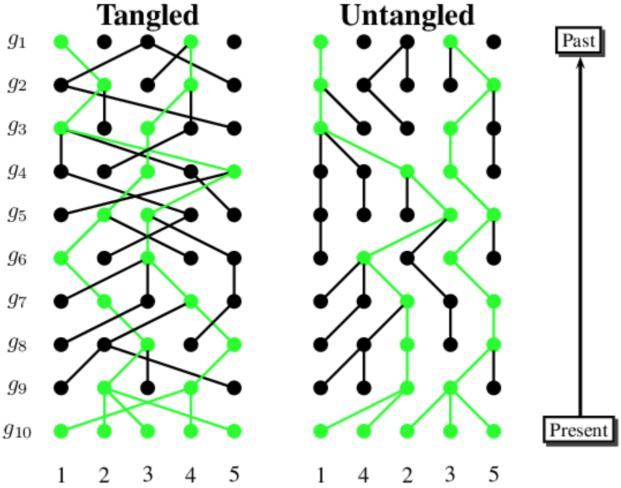
Genealogies are Random: 3 realisations when $N = 5$



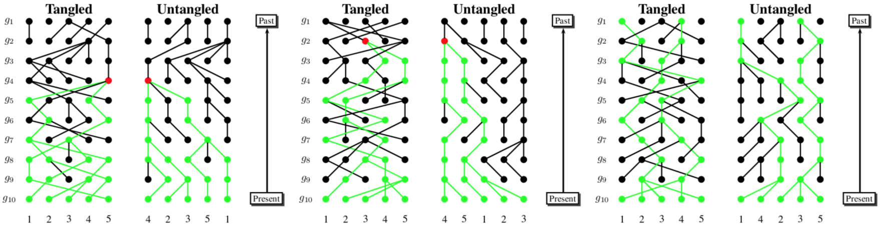
Genealogies are Random: 3 realisations when $N = 5$



Genealogies are Random: 3 realisations when $N = 5$

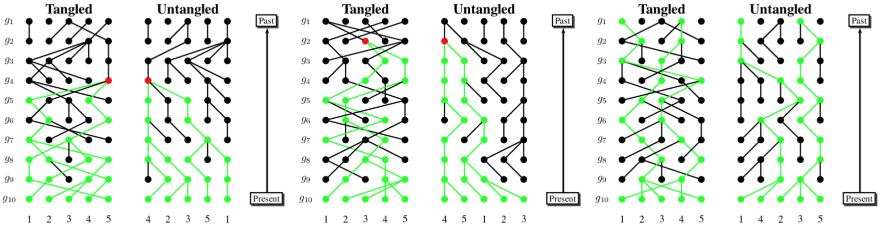


Genealogies are Random: 3 realisations when $N = 5$



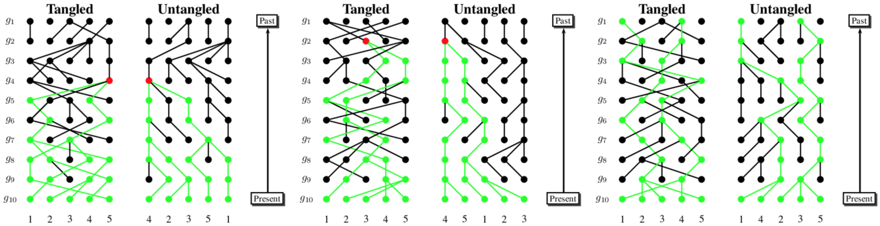
- Moreover, the population size N is typically large ($N > 10^4$)

Genealogies are Random: 3 realisations when $N = 5$



- Moreover, the population size N is typically large ($N > 10^4$)
- We are often interested in the genealogy of a random sample of size n that is much smaller than N ($n \approxeq 30$) from the present generation

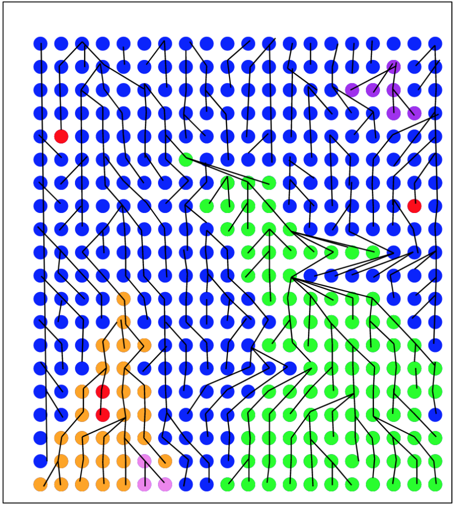
Genealogies are Random: 3 realisations when $N = 5$



- Moreover, the population size N is typically large ($N > 10^4$)
- We are often interested in the genealogy of a random sample of size n that is much smaller than N ($n \approxeq 30$) from the present generation
- Let us understand the idea of population and sample genealogies in a slightly larger population with $N = 20$ next

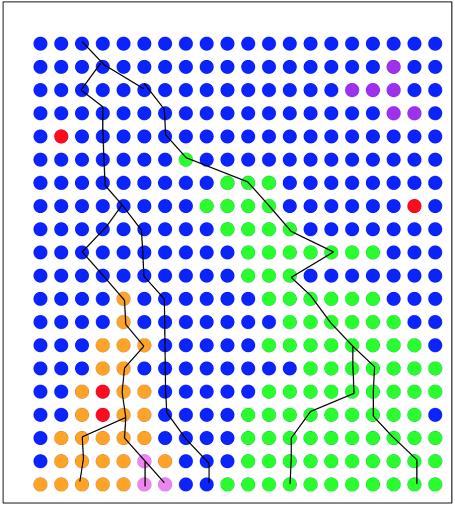
Population and Sample Genealogies when $N = 20$

- Instead of this population genealogy of all individuals in the present generation with population size $N = 20$,



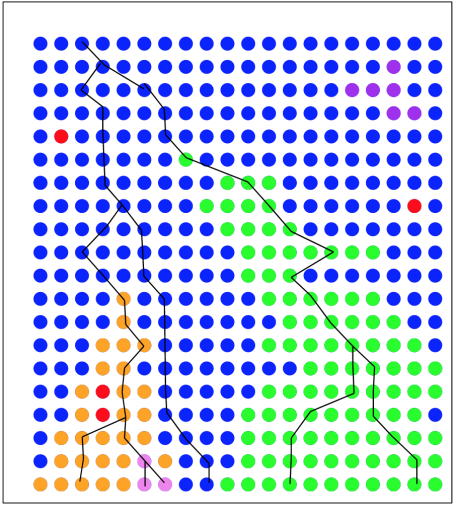
Population and Sample Genealogies when $N = 20$

- Instead of this population genealogy of all individuals in the present generation with population size $N = 20$,
- say, we are only interested in the genealogy of a random sample of size $n = 6$ from the present generation.



Population and Sample Genealogies when $N = 20$

- Instead of this population genealogy of all individuals in the present generation with population size $N = 20$,
- say, we are only interested in the genealogy of a random sample of size $n = 6$ from the present generation.
- Note that the sample genealogy is a sub-graph of the population genealogy and merely comes from only tracking the parents, parents' parents, and so on of the six sampled individuals.



Population and Sample Genealogies when $N = 20$

- Instead of this population genealogy of all individuals in the present generation with population size $N = 20$,
- say, we are only interested in the genealogy of a random sample of size $n = 6$ from the present generation.
- Note that the sample genealogy is a sub-graph of the population genealogy and merely comes from only tracking the parents, parents' parents, and so on of the six sampled individuals.
- Kingman's n-Coalescent Model, we study next, is a good approximation of the W-F sample genealogy and can shed light on some of its basic properties when n is smaller than N (Kingman, J. F. C. 1982. The coalescent. Stochastic Processes and Their Applications 13: 235-248.).

Mathematical Preliminaries 1 – a basic limit

Lemma

$$\lim_{N\to\infty}\left(1-\frac{c}{N}\right)^{Nt}=e^{-ct},\quad \textit{for constants } c \textit{ and } t>0.$$

Mathematical Preliminaries 1 – a basic limit

Lemma

$$\lim_{N\to\infty}\left(1-\frac{c}{N}\right)^{Nt}=e^{-ct},\quad \textit{for constants } c \textit{ and } t>0.$$

Proof.

$$\lim_{N\to\infty}\left(1-\tfrac{c}{N}\right)^{Nt}=\lim_{N\to\infty}\exp\left(Nt\ln\left(1-\tfrac{c}{N}\right)\right)$$

Mathematical Preliminaries 1 – a basic limit

Lemma

$$\lim_{N\to\infty}\left(1-\frac{c}{N}\right)^{Nt} = e^{-ct}, \quad \textit{for constants } c \textit{ and } t > 0.$$

Proof.

$$\lim_{N\to\infty}\left(1-\tfrac{c}{N}\right)^{Nt} = \lim_{N\to\infty}\exp\left(Nt\ln\left(1-\tfrac{c}{N}\right)\right) = \exp\left(\lim_{N\to\infty}Nt\ln\left(1-\tfrac{c}{N}\right)\right)$$

UPPSALA
UNIVERSITET

Mathematical Preliminaries 1 – a basic limit

Lemma

$$\lim_{N\to\infty}\left(1-\frac{c}{N}\right)^{Nt}=e^{-ct},\quad \textit{for constants } c \textit{ and } t>0.$$

Proof.

$$\lim_{N\to\infty}\left(1-\tfrac{c}{N}\right)^{Nt}=\lim_{N\to\infty}\exp\left(Nt\ln\left(1-\tfrac{c}{N}\right)\right)=\exp\left(\lim_{N\to\infty}Nt\ln\left(1-\tfrac{c}{N}\right)\right)=\exp\left(\lim_{N\to\infty}\frac{\ln\left(1-\frac{c}{N}\right)}{\frac{1}{Nt}}\right)$$

Mathematical Preliminaries 1 – a basic limit

Lemma

$$\lim_{N\to\infty}\left(1-\frac{c}{N}\right)^{Nt}=e^{-ct},\quad \textit{for constants } c \textit{ and } t>0.$$

Proof.

$$\lim_{N\to\infty}\left(1-\tfrac{c}{N}\right)^{Nt}=\lim_{N\to\infty}\exp\left(Nt\ln\left(1-\tfrac{c}{N}\right)\right)=\exp\left(\lim_{N\to\infty}Nt\ln\left(1-\tfrac{c}{N}\right)\right)=\exp\left(\lim_{N\to\infty}\frac{\ln\left(1-\frac{c}{N}\right)}{\frac{1}{Nt}}\right)$$

$$=\exp\left(\lim_{N\to\infty}\frac{\frac{d}{dN}\left(\ln\left(1-\frac{c}{N}\right)\right)}{\frac{d}{dN}\left(\frac{1}{Nt}\right)}\right),\quad \text{by L'Hopital's Rule to the indeterminate form } \tfrac{0}{0}$$

Mathematical Preliminaries 1 – a basic limit

Lemma

$$\lim_{N\to\infty}\left(1-\frac{c}{N}\right)^{Nt} = e^{-ct}, \quad \textit{for constants } c \textit{ and } t > 0.$$

Proof.

$$\lim_{N\to\infty}\left(1-\tfrac{c}{N}\right)^{Nt} = \lim_{N\to\infty}\exp\left(Nt\ln\left(1-\tfrac{c}{N}\right)\right) = \exp\left(\lim_{N\to\infty} Nt\ln\left(1-\tfrac{c}{N}\right)\right) = \exp\left(\lim_{N\to\infty}\frac{\ln\left(1-\frac{c}{N}\right)}{\frac{1}{Nt}}\right)$$

$$= \exp\left(\lim_{N\to\infty}\frac{\frac{d}{dN}\left(\ln\left(1-\frac{c}{N}\right)\right)}{\frac{d}{dN}\left(\frac{1}{Nt}\right)}\right), \quad \text{by L'Hopital's Rule to the indeterminate form } \tfrac{0}{0}$$

$$= \exp\left(\lim_{N\to\infty}\frac{\frac{c}{N^2-Nc}}{-\frac{1}{tN^2}}\right)$$

Mathematical Preliminaries 1 – a basic limit

Lemma

$$\lim_{N\to\infty}\left(1-\frac{c}{N}\right)^{Nt} = e^{-ct}, \quad \textit{for constants } c \textit{ and } t > 0.$$

Proof.

$$\lim_{N\to\infty}\left(1-\tfrac{c}{N}\right)^{Nt} = \lim_{N\to\infty}\exp\left(Nt\ln\left(1-\tfrac{c}{N}\right)\right) = \exp\left(\lim_{N\to\infty} Nt\ln\left(1-\tfrac{c}{N}\right)\right) = \exp\left(\lim_{N\to\infty}\frac{\ln\left(1-\frac{c}{N}\right)}{\frac{1}{Nt}}\right)$$

$$= \exp\left(\lim_{N\to\infty}\frac{\frac{d}{dN}\left(\ln\left(1-\frac{c}{N}\right)\right)}{\frac{d}{dN}\left(\frac{1}{Nt}\right)}\right), \quad \text{by L'Hopital's Rule to the indeterminate form } \tfrac{0}{0}$$

$$= \exp\left(\lim_{N\to\infty}\frac{\frac{c}{N^2-Nc}}{-\frac{1}{tN^2}}\right) = \exp\left(\lim_{N\to\infty}\frac{-ct}{\left(1-\frac{c}{N}\right)}\right)$$

Mathematical Preliminaries 1 – a basic limit

Lemma

$$\lim_{N\to\infty}\left(1-\frac{c}{N}\right)^{Nt} = e^{-ct}, \quad \textit{for constants } c \textit{ and } t > 0.$$

Proof.

$$\lim_{N\to\infty}\left(1-\tfrac{c}{N}\right)^{Nt} = \lim_{N\to\infty}\exp\left(Nt\ln\left(1-\tfrac{c}{N}\right)\right) = \exp\left(\lim_{N\to\infty} Nt\ln\left(1-\tfrac{c}{N}\right)\right) = \exp\left(\lim_{N\to\infty}\frac{\ln\left(1-\frac{c}{N}\right)}{\frac{1}{Nt}}\right)$$

$$= \exp\left(\lim_{N\to\infty}\frac{\frac{d}{dN}\left(\ln\left(1-\frac{c}{N}\right)\right)}{\frac{d}{dN}\left(\frac{1}{Nt}\right)}\right), \quad \text{by L'Hopital's Rule to the indeterminate form } \tfrac{0}{0}$$

$$= \exp\left(\lim_{N\to\infty}\frac{\frac{c}{N^2-Nc}}{-\frac{1}{tN^2}}\right) = \exp\left(\lim_{N\to\infty}\frac{-ct}{\left(1-\frac{c}{N}\right)}\right) = \exp\left(\frac{-ct}{\left(1-\lim\limits_{N\to\infty}\frac{c}{N}\right)}\right)$$

Mathematical Preliminaries 1 – a basic limit

Lemma

$$\lim_{N\to\infty}\left(1-\frac{c}{N}\right)^{Nt}=e^{-ct},\quad \textit{for constants } c \textit{ and } t>0.$$

Proof.

$$\lim_{N\to\infty}\left(1-\tfrac{c}{N}\right)^{Nt}=\lim_{N\to\infty}\exp\left(Nt\ln\left(1-\tfrac{c}{N}\right)\right)=\exp\left(\lim_{N\to\infty}Nt\ln\left(1-\tfrac{c}{N}\right)\right)=\exp\left(\lim_{N\to\infty}\frac{\ln\left(1-\frac{c}{N}\right)}{\frac{1}{Nt}}\right)$$

$$=\exp\left(\lim_{N\to\infty}\frac{\frac{d}{dN}\left(\ln\left(1-\frac{c}{N}\right)\right)}{\frac{d}{dN}\left(\frac{1}{Nt}\right)}\right),\quad \text{by L'Hopital's Rule to the indeterminate form } \tfrac{0}{0}$$

$$=\exp\left(\lim_{N\to\infty}\frac{\frac{c}{N^2-Nc}}{-\frac{1}{tN^2}}\right)=\exp\left(\lim_{N\to\infty}\frac{-ct}{\left(1-\frac{c}{N}\right)}\right)=\exp\left(\frac{-ct}{\left(1-\lim\limits_{N\to\infty}\frac{c}{N}\right)}\right)=\exp\left(\frac{-ct}{(1-0)}\right)$$

Mathematical Preliminaries 1 – a basic limit

Lemma

$$\lim_{N\to\infty}\left(1-\frac{c}{N}\right)^{Nt} = e^{-ct}, \quad \textit{for constants } c \textit{ and } t > 0.$$

Proof.

$$\begin{aligned}
\lim_{N\to\infty}\left(1-\tfrac{c}{N}\right)^{Nt} &= \lim_{N\to\infty}\exp\left(Nt\ln\left(1-\tfrac{c}{N}\right)\right) = \exp\left(\lim_{N\to\infty} Nt\ln\left(1-\tfrac{c}{N}\right)\right) = \exp\left(\lim_{N\to\infty}\frac{\ln\left(1-\frac{c}{N}\right)}{\frac{1}{Nt}}\right)\\
&= \exp\left(\lim_{N\to\infty}\frac{\frac{d}{dN}\left(\ln\left(1-\frac{c}{N}\right)\right)}{\frac{d}{dN}\left(\frac{1}{Nt}\right)}\right), \quad \text{by L'Hopital's Rule to the indeterminate form } \tfrac{0}{0}\\
&= \exp\left(\lim_{N\to\infty}\frac{\frac{c}{N^2-Nc}}{-\frac{1}{tN^2}}\right) = \exp\left(\lim_{N\to\infty}\frac{-ct}{\left(1-\frac{c}{N}\right)}\right) = \exp\left(\frac{-ct}{\left(1-\lim\limits_{N\to\infty}\frac{c}{N}\right)}\right) = \exp\left(\frac{-ct}{(1-0)}\right)\\
&= e^{-ct}
\end{aligned}$$

$\square$

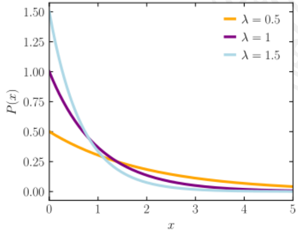
Mathematical Preliminaries 2 – $\mathrm{Exponential}(\lambda)$ R.V.

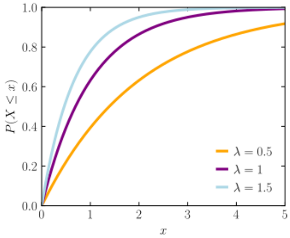
The probability density function (PDF) of $\mathrm{Exponential}(\lambda)$ R.V. with rate parameter $\lambda > 0$ is

$$f(x;\lambda) = \begin{cases} \lambda e^{-\lambda x} & x \geq 0, \\ 0 & x < 0. \end{cases}$$

And its cumulative distribution function (CDF) is

$$F(x;\lambda) = \begin{cases} 1 - e^{-\lambda x} & x \geq 0, \\ 0 & x < 0. \end{cases}$$





Mathematical Preliminaries 2 – $\mathrm{Exponential}(\lambda)$ R.V.

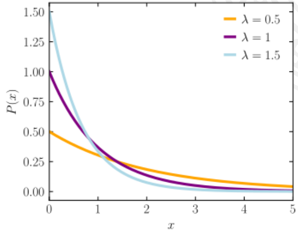
The probability density function (PDF) of $\mathrm{Exponential}(\lambda)$ R.V. with rate parameter $\lambda > 0$ is

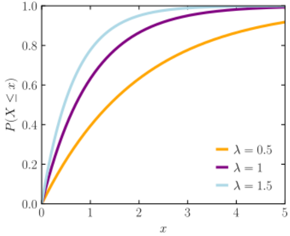
$$f(x;\lambda) = \begin{cases} \lambda e^{-\lambda x} & x \geq 0, \\ 0 & x < 0. \end{cases}$$

And its cumulative distribution function (CDF) is

$$F(x;\lambda) = \begin{cases} 1 - e^{-\lambda x} & x \geq 0, \\ 0 & x < 0. \end{cases}$$

Its Mean or Expectation is $1/\lambda$ and Variance is $1/\lambda^2$.





Mathematical Preliminaries 2 – $\text{Exponential}(\lambda)$ R.V.

The probability density function (PDF) of $\text{Exponential}(\lambda)$ R.V. with rate parameter $\lambda > 0$ is

$$f(x;\lambda) = \begin{cases} \lambda e^{-\lambda x} & x \geq 0, \\ 0 & x < 0. \end{cases}$$

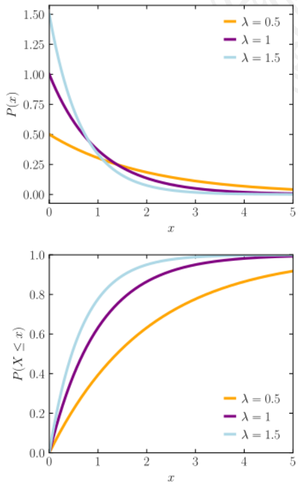
And its cumulative distribution function (CDF) is

$$F(x;\lambda) = \begin{cases} 1 - e^{-\lambda x} & x \geq 0, \\ 0 & x < 0. \end{cases}$$

Its Mean or Expectation is $1/\lambda$ and Variance is $1/\lambda^2$.

Lemma (Competing Exponentials)

Let $X_1, X_2, \ldots, X_n$ be independent exponentially distributed random variables with rate parameters $\lambda_1, \lambda_2, \ldots, \lambda_n$. Then $\min\{X_1, \ldots, X_n\}$ *is also exponentially distributed, with rate parameter* $\lambda = \lambda_1 + \cdots + \lambda_n$.



Kingman's *n*-Coalescent: sample of size 2

- $\Pr(\text{2 individuals have the same parent or coalesce in the previous generation}) = \frac{1}{N}$

Kingman's *n*-Coalescent: sample of size 2

- $\Pr(\text{2 individuals have the same parent or coalesce in the previous generation}) = \frac{1}{N}$
- $\Pr(\text{2 individuals do not coalesce for more than } g \text{ generations}) = \left(1 - \frac{1}{N}\right)^g$

Kingman's *n*-Coalescent: sample of size 2

- $\Pr(\text{2 individuals have the same parent or coalesce in the previous generation}) = \frac{1}{N}$
- $\Pr(\text{2 individuals do not coalesce for more than } g \text{ generations}) = \left(1 - \frac{1}{N}\right)^g$
- Let rescaled continuous time t be g in units of N generations. Then,

Kingman's *n*-Coalescent: sample of size 2

- $\Pr(\text{2 individuals have the same parent or coalesce in the previous generation}) = \frac{1}{N}$
- $\Pr(\text{2 individuals do not coalesce for more than } g \text{ generations}) = \left(1 - \frac{1}{N}\right)^{g}$
- Let rescaled continuous time t be g in units of N generations. Then,
- $\Pr(\text{2 individuals do not coalesce by rescaled time } t) = \left(1 - \frac{1}{N}\right)^{\lfloor Nt \rfloor} \xrightarrow[N\to\infty]{} e^{-t}$

Kingman's *n*-Coalescent: sample of size 2

- $\Pr(\text{2 individuals have the same parent or coalesce in the previous generation}) = \frac{1}{N}$
- $\Pr(\text{2 individuals do not coalesce for more than } g \text{ generations}) = \left(1 - \frac{1}{N}\right)^{g}$
- Let rescaled continuous time t be g in units of N generations. Then,
- $\Pr(\text{2 individuals do not coalesce by rescaled time } t) = \left(1 - \frac{1}{N}\right)^{\lfloor Nt \rfloor} \xrightarrow[N\to\infty]{} e^{-t}$
- $\Pr(\text{2 individuals do coalesce by rescaled time } t) = 1 - e^{-t}$

UPPSALA
UNIVERSITET

Kingman's n-Coalescent: sample of size 2

- $\Pr(\text{2 individuals have the same parent or coalesce in the previous generation}) = \frac{1}{N}$
- $\Pr(\text{2 individuals do not coalesce for more than } g \text{ generations}) = \left(1 - \frac{1}{N}\right)^g$
- Let rescaled continuous time t be g in units of N generations. Then,
- $\Pr(\text{2 individuals do not coalesce by rescaled time } t) = \left(1 - \frac{1}{N}\right)^{\lfloor Nt \rfloor} \underset{N\to\infty}{\longrightarrow} e^{-t}$
- $\Pr(\text{2 individuals do coalesce by rescaled time } t) = 1 - e^{-t}$
- Note that $1 - e^{-t}$ is the distribution function of the $\mathrm{Exponential}(1)$, the Exponential random variable with rate parameter 1

Kingman's *n*-Coalescent: sample of size 2

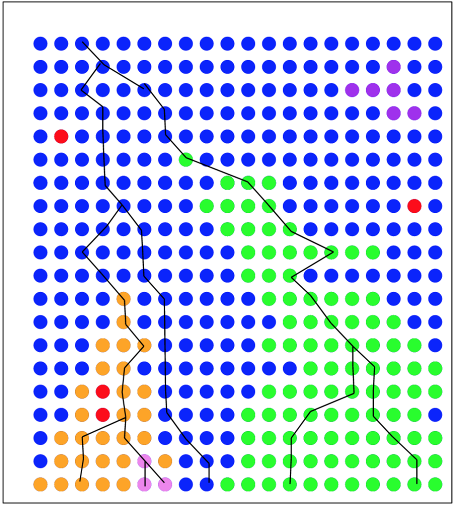
- $\Pr(\text{2 individuals have the same parent or coalesce in the previous generation}) = \frac{1}{N}$
- $\Pr(\text{2 individuals do not coalesce for more than } g \text{ generations}) = \left(1 - \frac{1}{N}\right)^g$
- Let rescaled continuous time t be g in units of N generations. Then,
- $\Pr(\text{2 individuals do not coalesce by rescaled time } t) = \left(1 - \frac{1}{N}\right)^{\lfloor Nt \rfloor} \underset{N\to\infty}{\longrightarrow} e^{-t}$
- $\Pr(\text{2 individuals do coalesce by rescaled time } t) = 1 - e^{-t}$
- Note that $1 - e^{-t}$ is the distribution function of the $\mathrm{Exponential}(1)$, the Exponential random variable with rate parameter 1
- Thus the time for two individuals to coalesce is given by $\mathrm{Exponential}(1)$ random variable with density function e^{-t}

Kingman's *n*-Coalescent: sample of size 2

- $\Pr(\text{2 individuals have the same parent or coalesce in the previous generation}) = \frac{1}{N}$
- $\Pr(\text{2 individuals do not coalesce for more than } g \text{ generations}) = \left(1 - \frac{1}{N}\right)^g$
- Let rescaled continuous time t be g in units of N generations. Then,
- $\Pr(\text{2 individuals do not coalesce by rescaled time } t) = \left(1 - \frac{1}{N}\right)^{\lfloor Nt \rfloor} \xrightarrow[N\to\infty]{} e^{-t}$
- $\Pr(\text{2 individuals do coalesce by rescaled time } t) = 1 - e^{-t}$
- Note that $1 - e^{-t}$ is the distribution function of the $\mathrm{Exponential}(1)$, the Exponential random variable with rate parameter 1
- Thus the time for two individuals to coalesce is given by $\mathrm{Exponential}(1)$ random variable with density function e^{-t}
- This is the main idea of Kingman's approximation for the coalescence time of a random sample of two individuals from a population whose size N gets large in the limit.

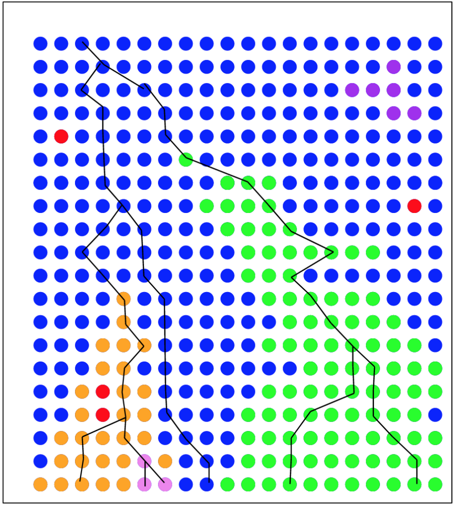
Kingman's n-Coalescent: sample of size $2 \leq i \leq n$

- Consider the genealogy of a random sample of size $n = 6$ from the present generation to fix ideas and let $2 \leq i \leq n$



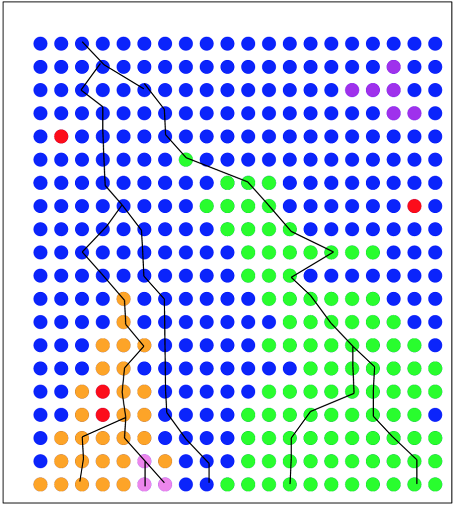
Kingman's n-Coalescent: sample of size $2 \leq i \leq n$

- Consider the genealogy of a random sample of size $n = 6$ from the present generation to fix ideas and let $2 \leq i \leq n$
- When there are i ancestral lineages of the sample that have not coalesced yet, there are $\binom{i}{2}$ pairs of lineages, each of which coalesce independently according to $\mathrm{Exponential}(1)$ random variable.



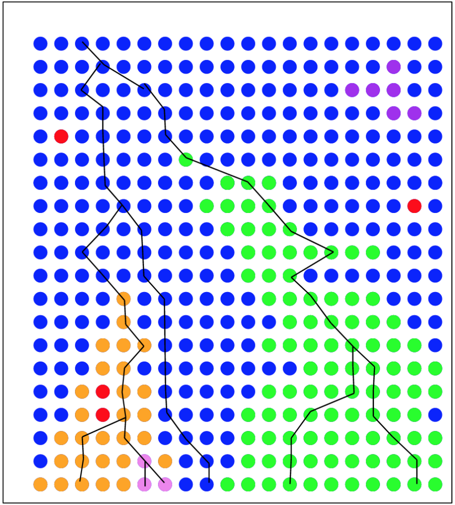
Kingman's n-Coalescent: sample of size $2 \leq i \leq n$

- Consider the genealogy of a random sample of size $n = 6$ from the present generation to fix ideas and let $2 \leq i \leq n$
- When there are i ancestral lineages of the sample that have not coalesced yet, there are $\binom{i}{2}$ pairs of lineages, each of which coalesce independently according to $\mathrm{Exponential}(1)$ random variable.
- Thus, any one of the pairs coalescing is given by the minimum of the $\binom{i}{2}$ many competing independent $\mathrm{Exponential}(1)$ random variables – which is given by $\mathrm{Exponential}(\binom{i}{2})$ random variable whose rate is obtained by summing the rate 1 of each pair.



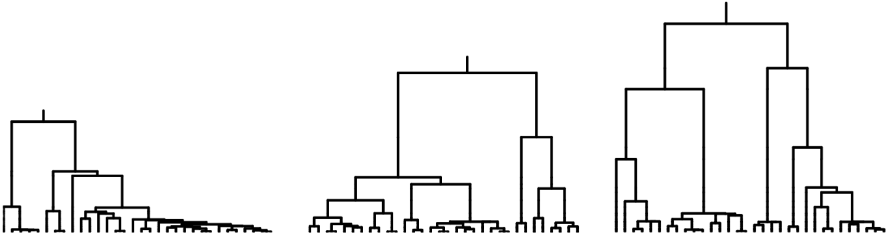
Kingman’s n-Coalescent: sample of size $2 \leq i \leq n$

- Consider the genealogy of a random sample of size $n = 6$ from the present generation to fix ideas and let $2 \leq i \leq n$
- When there are i ancestral lineages of the sample that have not coalesced yet, there are $\binom{i}{2}$ pairs of lineages, each of which coalesce independently according to $\mathrm{Exponential}(1)$ random variable.
- Thus, any one of the pairs coalescing is given by the minimum of the $\binom{i}{2}$ many competing independent $\mathrm{Exponential}(1)$ random variables – which is given by $\mathrm{Exponential}(\binom{i}{2})$ random variable whose rate is obtained by summing the rate 1 of each pair.
- This already gives us a simulation algorithm of sample genealogies under Kingman’s n-Coalescent Model.



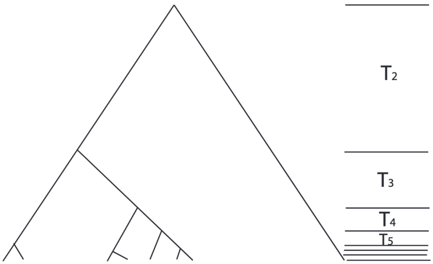
Kingman’s *n*-Coalescent: Simulating Genealogies

- Draw a sample from $\mathrm{Exponential}(\binom{n}{2})$ random variable.
- Pick two lineages at random to coalescence.
- Decrease n by 1
- If $n = 1$, stop; else, repeat these steps.



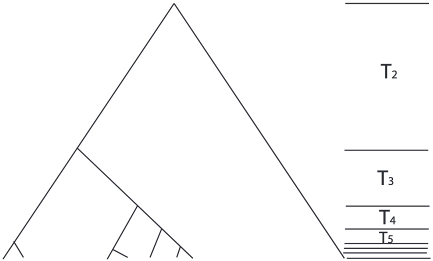
Kingman's *n*-Coalescent: Genealogical Times

- Let T_i be the time in the history of the sample when there were exactly i ancestral lineages



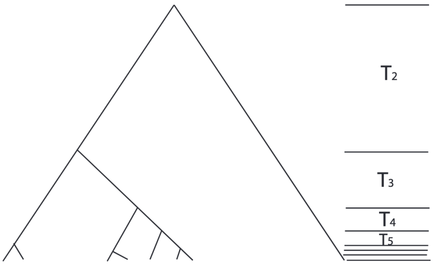
Kingman's *n*-Coalescent: Genealogical Times

- Let T_i be the time in the history of the sample when there were exactly i ancestral lineages
- Time to the most recent common ancestor (MRCA) of all lineages sampled = $T_{\mathrm{MRCA}} = \sum_{i=2}^{n} T_i$



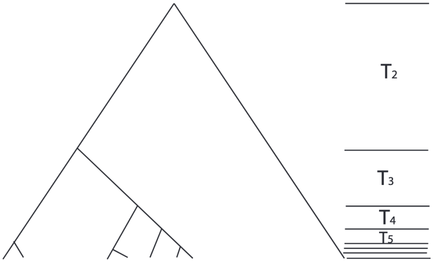
Kingman's *n*-Coalescent: Genealogical Times

- Let T_i be the time in the history of the sample when there were exactly i ancestral lineages
- Time to the most recent common ancestor (MRCA) of all lineages sampled = $T_{\mathrm{MRCA}} = \sum_{i=2}^{n} T_i$
- Total time represented by the genealogy = $T_{\mathrm{TOT}} = \sum_{i=2}^{n} iT_i$



Kingman's *n*-Coalescent: Genealogical Times

- Let T_i be the time in the history of the sample when there were exactly i ancestral lineages
- Time to the most recent common ancestor (MRCA) of all lineages sampled = $T_{\mathrm{MRCA}} = \sum_{i=2}^{n} T_i$
- Total time represented by the genealogy = $T_{\mathrm{TOT}} = \sum_{i=2}^{n} iT_i$
- We are interested in properties of T_{MRCA} and T_{TOT}.



Kingman's *n*-Coalescent: Means of T_{MRCA} and T_{TOT}

- The mean or expected time to the MRCA is

$$E\left(T_{\mathrm{MRCA}}\right) = E\left(\sum_{i=2}^{n} T_i\right) = \sum_{i=2}^{n} E(T_i) = \sum_{i=2}^{n} \frac{2}{i(i-1)} = 2\left(1 - \frac{1}{n}\right)$$

Kingman's n-Coalescent: Means of T_{MRCA} and T_{TOT}

- The mean or expected time to the MRCA is

$$E(T_{\text{MRCA}}) = E\left(\sum_{i=2}^{n} T_i\right) = \sum_{i=2}^{n} E(T_i) = \sum_{i=2}^{n} \frac{2}{i(i-1)} = 2\left(1 - \frac{1}{n}\right)$$

- The mean or expected Total time represented by the genealogy is

$$E(T_{\text{TOT}}) = E\left(\sum_{i=2}^{n} iT_i\right) = \sum_{i=2}^{n} iE(T_i) = \sum_{i=2}^{n} \frac{2i}{i(i-1)} = \sum_{i=1}^{n-1} \frac{2}{i} = 2\sum_{i=1}^{n-1} \frac{1}{i}$$

UPPSALA
UNIVERSITET

Kingman's n-Coalescent: Means of T_{MRCA} and T_{TOT}

- The mean or expected time to the MRCA is

$$E\left(T_{\mathrm{MRCA}}\right) = E\left(\sum_{i=2}^{n} T_i\right) = \sum_{i=2}^{n} E(T_i) = \sum_{i=2}^{n} \frac{2}{i(i-1)} = 2\left(1 - \frac{1}{n}\right)$$

- The mean or expected Total time represented by the genealogy is

$$E\left(T_{\mathrm{TOT}}\right) = E\left(\sum_{i=2}^{n} iT_i\right) = \sum_{i=2}^{n} iE(T_i) = \sum_{i=2}^{n} \frac{2i}{i(i-1)} = \sum_{i=1}^{n-1} \frac{2}{i} = 2\sum_{i=1}^{n-1} \frac{1}{i}$$

- The Variance as well as the probability densities of T_{MRCA} and T_{TOT} are also known.

Kingman's n-Coalescent: Means of T_{MRCA} and T_{TOT}

- The mean or expected time to the MRCA is

$$E\left(T_{\mathrm{MRCA}}\right) = E\left(\sum_{i=2}^{n} T_i\right) = \sum_{i=2}^{n} E(T_i) = \sum_{i=2}^{n} \frac{2}{i(i-1)} = 2\left(1 - \frac{1}{n}\right)$$

- The mean or expected Total time represented by the genealogy is

$$E\left(T_{\mathrm{TOT}}\right) = E\left(\sum_{i=2}^{n} iT_i\right) = \sum_{i=2}^{n} iE(T_i) = \sum_{i=2}^{n} \frac{2i}{i(i-1)} = \sum_{i=1}^{n-1} \frac{2}{i} = 2\sum_{i=1}^{n-1} \frac{1}{i}$$

- The Variance as well as the probability densities of T_{MRCA} and T_{TOT} are also known.

Summary

Kingman's n-Coalescent model is a much simpler approximation to the sample genealogy of the the Wright-Fisher model, provided we have a random sample of size n that is much smaller than the population size N, by rescaling time in units of N generations.

Extensions to Coalescent Models

All these extensions are possible and routinely used in genomic sciences today:

- Allow DNA sequences obtained from the sampled individuals to coalesce through the genealogies by also modeling DNA mutations to occur at given mutation rates
- Allow models to have demographic changes through time
- Allow sub-populations structure and migration between them
- Allow some DNA sequences to recombine and be trackable into both parents
- Allow samples from populations of different species
- etc.
- And, finally use the extended coalescent models for inference (testing and prediction) based on sampled DNA sequences

Coalescent Models with Extensions in Action

Article | Open Access | Published: 07 April 2021

Initial Upper Palaeolithic humans in Europe had recent Neanderthal ancestry

Mateja Hajdinjak, Fabrizio Mafessoni, Laurits Skov, Benjamin Vernot, Alexander Hübner, Qiaomei Fu, Elena Essel, Sarah Nagel, Birgit Nickel, Julia Richter, Oana Teodora Moldovan, Silviu Constantin, Elena Endarova, Nikolay Zahariev, Rosen Spasov, Frido Welker, Geoff M. Smith, Virginie Sinet-Mathiot, Lindsey Paskulin, Helen Fewlass, Sahra Talamo, Zeljko Rezek, Svoboda Sirakova, Nikolay Sirakov, ... Svante Pääbo

Nature **592**, 253–257 (2021) | Cite this article

72k Accesses | **42** Citations | **1276** Altmetric | Metrics

This article has been updated

Abstract

Modern humans appeared in Europe by at least 45,000 years ago[1,2,3,4,5], but the extent of their interactions with Neanderthals, who disappeared by about 40,000 years ago[6], and their relationship to the broader expansion of modern humans outside Africa are poorly understood. Here we present genome-wide data from three individuals dated to between

UPPSALA
UNIVERSITET

Outlook: Current Research Perspectives

- Empirical Pedigree Processes
- Diploid Recombining Ancestries

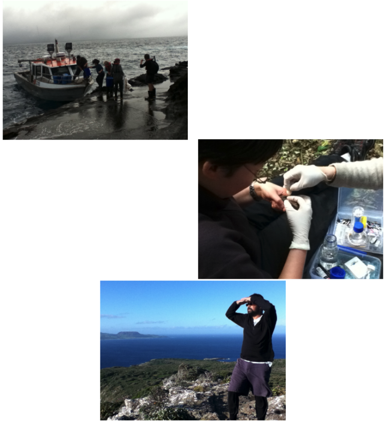
UPPSALA
UNIVERSITET

Empirical Pedigree Processes

Inspired by Chatham Island Black Robins.

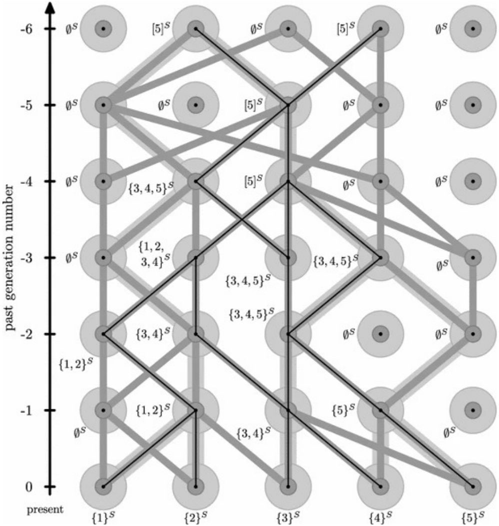
Empirical Pedigree Processes

Insights from field work in November 2011



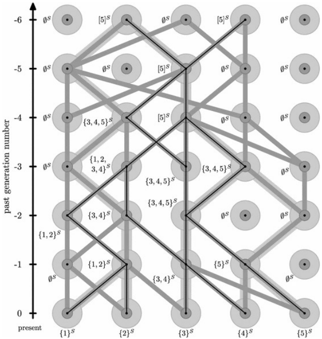
Inspired by Chatham Island Black Robins.

Diploid Recombining Ancestral Process



UPPSALA
UNIVERSITET

Diploid Recombining Ancestral Process – TMRCA



Theorem (Sainudiin-Thatte-Veber)

Let T_n denote the number of generations, counting back in time from the present, to an MRCA of all present-day individuals. Then for every $\varepsilon > 0$

$$\lim_{n\to\infty} \mathbb{P}\{(1-\varepsilon)C(r)\ln n \le T_n \le (1+\varepsilon)C(r)\ln n\} = 1,,$$

$$C(r) := \frac{1}{\ln(1+r)} - \frac{1}{\ln(1-r)}.$$

Want to learn more?

I recommend lectures by:

- Mark Beaumont `https://gtpb.github.io/PGDH19/assets/coal_course/lecture1.pdf`
- John Wakeley `https://www.youtube.com/watch?v=lo864lPyszg`

Classical Introductory Textbook by:

- Coalescent Theory: An Introduction by Wakeley. `https://www.macmillanlearning.com/college/us/product/Coalescent-Theory/p/0974707759`