

The Wright-Fisher Model & Standard Coalescent

BIOL 1435

February 7, 2023

Overview

1. Quiz Review
2. Wright Fisher Model
3. Standard Coalescent Model
4. Derivations

Overview

1. Quiz Review

2. Wright Fisher Model

3. Standard Coalescent Model

4. Derivations

Segregating sites (S)

Definition

A segregating site is a site that is polymorphic in the data—i.e., there are multiple alleles observed.

Segregating sites (S)

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

Segregating sites (S)

$$\begin{bmatrix} 0 & 0 & 0 \\ 0 & 1 & 0 \\ 1 & 1 & 0 \\ 1 & 1 & 1 \end{bmatrix} \rightarrow \begin{bmatrix} 0 \\ 1 \\ 1 \\ 0 \end{bmatrix} \rightarrow S = 2$$

Gene diversity (h & H)

Definition

Gene diversity is the probability that two random DNA sequences are different.

Equation

$$h = 1 - \left(p^2 + (1 - p)^2 \right) \quad (1)$$

Where p is the frequency of the derived/alternative allele at a given site.

$$H = \frac{1}{L} \sum_{j=1}^L h_j \quad (2)$$

Where h_j is the gene diversity for site j and L is to the total number of sites.

Gene diversity (h & H)

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

Gene diversity (h & H)

$$\begin{bmatrix} 0 & 1 & 1 & 0 \\ 1 & 1 & 1 & 0 \\ 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 1 \end{bmatrix} \rightarrow \begin{bmatrix} 1 - (2^2/4^2 + 2^2/4^2) \\ 1 - (3^2/4^2 + 1^2/4^2) \\ 1 - (1^2/4^2 + 3^2/4^2) \\ 1 - (2^2/4^2 + 2^2/4^2) \end{bmatrix}$$

Gene diversity (h & H)

$$\begin{bmatrix} 0 & 1 & 1 & 0 \\ 1 & 1 & 1 & 0 \\ 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 1 \end{bmatrix} \rightarrow \begin{bmatrix} 1 - (2^2/4^2 + 2^2/4^2) \\ 1 - (3^2/4^2 + 1^2/4^2) \\ 1 - (1^2/4^2 + 3^2/4^2) \\ 1 - (2^2/4^2 + 2^2/4^2) \end{bmatrix} \rightarrow h_j = \begin{bmatrix} 8/16 \\ 6/16 \\ 6/16 \\ 8/16 \end{bmatrix}$$

Gene diversity (h & H)

$$\begin{bmatrix} 0 & 1 & 1 & 0 \\ 1 & 1 & 1 & 0 \\ 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 1 \end{bmatrix} \rightarrow \begin{bmatrix} 1 - (2^2/4^2 + 2^2/4^2) \\ 1 - (3^2/4^2 + 1^2/4^2) \\ 1 - (1^2/4^2 + 3^2/4^2) \\ 1 - (2^2/4^2 + 2^2/4^2) \end{bmatrix} \rightarrow h_j = \begin{bmatrix} 8/16 \\ 6/16 \\ 6/16 \\ 8/16 \end{bmatrix}$$

$$H = \frac{28}{16} \times \frac{1}{4} = \frac{7}{16}$$

Nucleotide diversity (Π & π)

Definition

Nucleotide diversity is the average number of pairwise differences between genotypes drawn from the same population.

Nucleotide diversity (Π & π)

Equation

$$\Pi = \sum_{i < j} k_{ij} \quad (3)$$

Where k_{ij} is the number of nucleotide differences between the i^{th} and j^{th} sample at a given site.

$$\pi = \frac{1}{\binom{n}{2}} \sum_{j=1}^L \Pi_j \quad (4)$$

Where Π_j is the nucleotide diversity for site j , L is the total number of sites in the genotype matrix, and $\binom{n}{2}$ is number of unique comparisons between n samples in the genotype matrix.

Nucleotide diversity (Π & π)

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

Nucleotide diversity (Π & π)

$$\begin{bmatrix} 0 & 1 & 1 & 0 \\ 1 & 1 & 1 & 0 \\ 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 1 \end{bmatrix} \rightarrow \begin{bmatrix} 2 + 1 + 1 \\ 1 + 1 + 1 \\ 3 + 0 + 0 \\ 2 + 1 + 1 \end{bmatrix} \rightarrow \Pi_j = \begin{bmatrix} 4 \\ 3 \\ 3 \\ 4 \end{bmatrix}$$

Nucleotide diversity (Π & π)

$$\begin{bmatrix} 0 & 1 & 1 & 0 \\ 1 & 1 & 1 & 0 \\ 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 1 \end{bmatrix} \rightarrow \begin{bmatrix} 2 + 1 + 1 \\ 1 + 1 + 1 \\ 3 + 0 + 0 \\ 2 + 1 + 1 \end{bmatrix} \rightarrow \Pi_j = \begin{bmatrix} 4 \\ 3 \\ 3 \\ 4 \end{bmatrix}$$

$$\pi = 14 \div \binom{4}{2} = 14 \div \frac{4(4-1)}{2} = \frac{14}{6}$$

Average nucleotide diversity in the presence of missing data (π_{pixy})

Equation

$$Numerator_{pixy} = \frac{1}{L_{called}} \sum_{j=1}^{L_{called}} \Pi_j \quad (5)$$

$$Denominator_{pixy} = \frac{1}{L_{called}} \sum_{j=1}^{L_{called}} \binom{n_{called}}{2}_j \quad (6)$$

$$\pi_{pixy} = \frac{Numerator_{pixy}}{Denominator_{pixy}} \quad (7)$$

Where Π_j is the nucleotide diversity among called genotypes for site j , L_{called} is to the total number of sites in the genotype matrix that have at least one called genotype, and $\binom{n_{called}}{2}_j$ is number of unique comparisons between n_{called} samples with genotype information at site j .

π_{pixy}

$$\begin{bmatrix} 0 & 0 & 0 & 0 & 1 \\ 1 & - & 0 & 1 & 1 \\ 0 & 0 & 0 & 0 & 0 \\ 1 & - & - & 1 & 0 \\ 0 & 0 & - & 1 & 1 \end{bmatrix}$$

π_{pixy}

$$\begin{bmatrix} 0 & 0 & 0 & 0 & 1 \\ 1 & - & 0 & 1 & 1 \\ 0 & 0 & 0 & 0 & 0 \\ 1 & - & - & 1 & 0 \\ 0 & 0 & - & 1 & 1 \end{bmatrix} \rightarrow \begin{bmatrix} 1 + 1 + 1 + 1 \\ 1 + 2 + 0 \\ 0 + 0 + 0 + 0 \\ 1 + 1 \\ 2 + 2 + 0 \end{bmatrix}$$

$$\begin{bmatrix} 0 & 0 & 0 & 0 & 1 \\ 1 & - & 0 & 1 & 1 \\ 0 & 0 & 0 & 0 & 0 \\ 1 & - & - & 1 & 0 \\ 0 & 0 & - & 1 & 1 \end{bmatrix} \rightarrow \begin{bmatrix} 1 + 1 + 1 + 1 \\ 1 + 2 + 0 \\ 0 + 0 + 0 + 0 \\ 1 + 1 \\ 2 + 2 + 0 \end{bmatrix} \rightarrow \Pi_j = \begin{bmatrix} 4 \\ 3 \\ 0 \\ 2 \\ 4 \end{bmatrix}$$

$$\begin{bmatrix} 0 & 0 & 0 & 0 & 1 \\ 1 & - & 0 & 1 & 1 \\ 0 & 0 & 0 & 0 & 0 \\ 1 & - & - & 1 & 0 \\ 0 & 0 & - & 1 & 1 \end{bmatrix} \rightarrow \begin{bmatrix} 1 + 1 + 1 + 1 \\ 1 + 2 + 0 \\ 0 + 0 + 0 + 0 \\ 1 + 1 \\ 2 + 2 + 0 \end{bmatrix} \rightarrow \Pi_j = \begin{bmatrix} 4 \\ 3 \\ 0 \\ 2 \\ 4 \end{bmatrix} \rightarrow \binom{n}{2}_j = \begin{bmatrix} 5C2 = 10 \\ 4C2 = 6 \\ 5C2 = 10 \\ 3C2 = 3 \\ 4C2 = 6 \end{bmatrix}$$

π_{pixy}

$$\begin{bmatrix} 0 & 0 & 0 & 0 & 1 \\ 1 & - & 0 & 1 & 1 \\ 0 & 0 & 0 & 0 & 0 \\ 1 & - & - & 1 & 0 \\ 0 & 0 & - & 1 & 1 \end{bmatrix} \rightarrow \begin{bmatrix} 1 + 1 + 1 + 1 \\ 1 + 2 + 0 \\ 0 + 0 + 0 + 0 \\ 1 + 1 \\ 2 + 2 + 0 \end{bmatrix} \rightarrow \Pi_j = \begin{bmatrix} 4 \\ 3 \\ 0 \\ 2 \\ 4 \end{bmatrix} \rightarrow \binom{n}{2}_j = \begin{bmatrix} 5C2 = 10 \\ 4C2 = 6 \\ 5C2 = 10 \\ 3C2 = 3 \\ 4C2 = 6 \end{bmatrix}$$

$$Numerator_{pixy} = \frac{13}{5}$$

$$Denominator_{pixy} = \frac{35}{5}$$

$$\pi_{pixy} = \frac{13}{5} \div \frac{35}{5} = \frac{13}{35}$$

Overview

1. Quiz Review

2. Wright Fisher Model

3. Standard Coalescent Model

4. Derivations

Modeling evolution

Q: Why do we need to model evolution?

Modeling evolution

A: To assess departures from neutrality.

Assumptions of the Wright Fisher model

Assumptions of the Wright Fisher model

- Panmictic population

Assumptions of the Wright Fisher model

- Panmictic population
- Constant population size of N
- Total of $2N$ allele copies

Assumptions of the Wright Fisher model

- Panmictic population
- Constant population size of N
- Total of $2N$ allele copies
- Discrete time process with non-overlapping generations
- All mutations are neutral

WF population ($N = 5$)

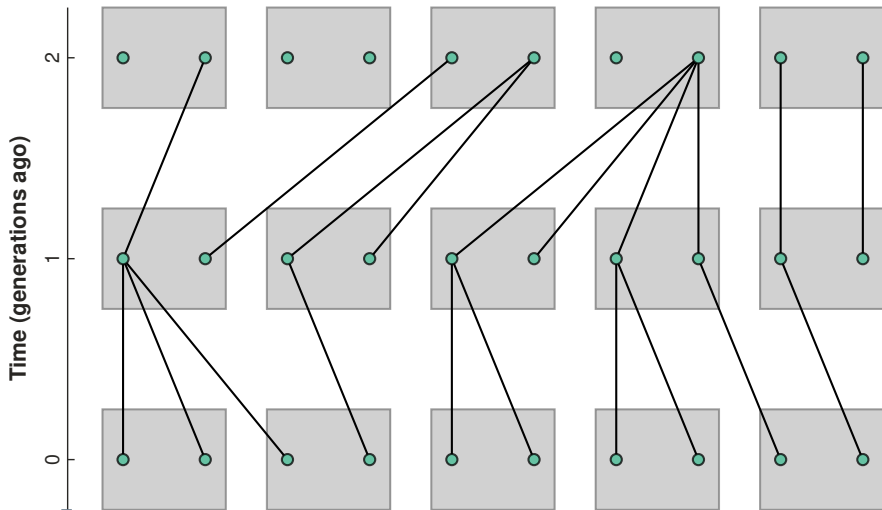


Simulating under the WF model

Simulating Reproduction

Sample with replacement an allelic copy (i) in the current generation (t) with a probability of $\frac{1}{2N}$ to produce an offspring in the next generation ($t + 1$), until there are $2N$ allelic copies in the next generation.

WF population ($N = 5$)



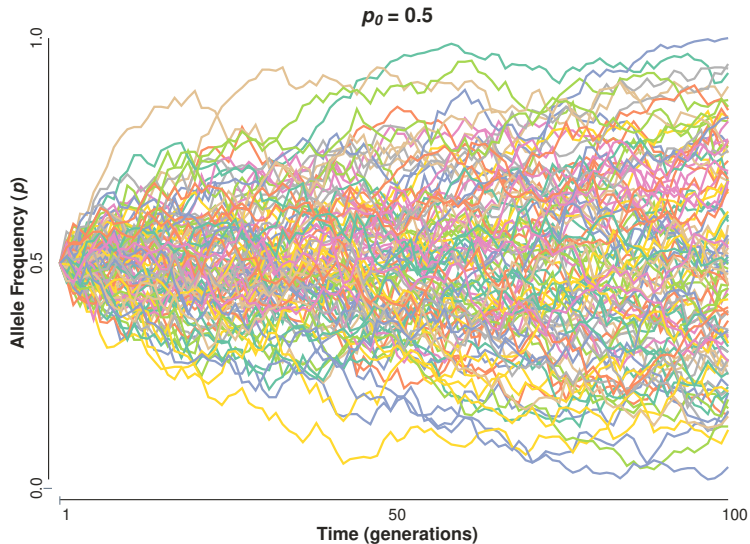
WF model & genetic drift

Definition

$$Pr(j) = \binom{2N}{j} \left(\frac{j}{2N}\right)^j \left(\frac{2N-j}{2N}\right)^{2N-j} \quad (8)$$

Where j represents the number of allelic copies of a particular allele.

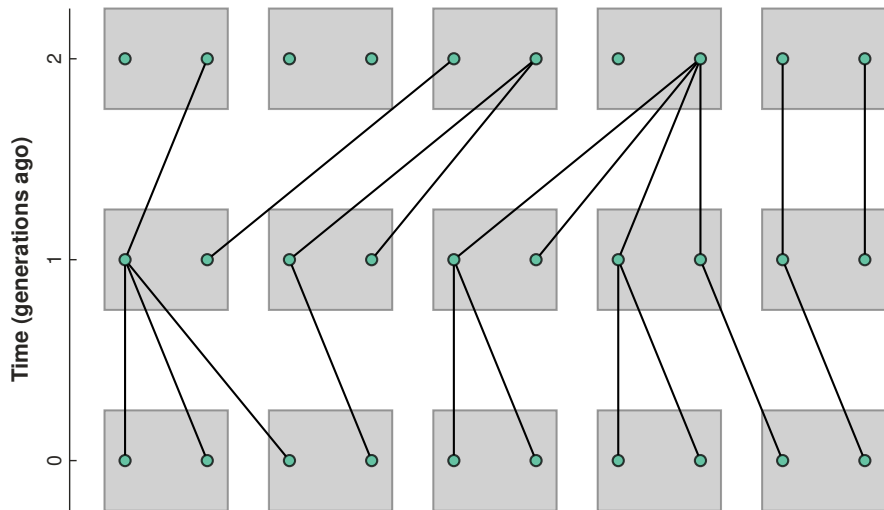
WF model & genetic drift



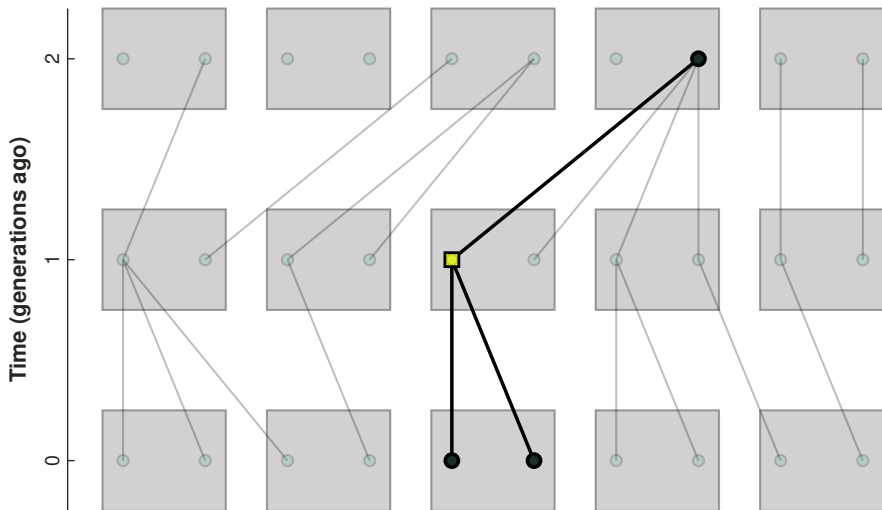
Overview

1. Quiz Review
2. Wright Fisher Model
- 3. Standard Coalescent Model**
4. Derivations

The standard coalescent



Finding the same parent = coalescence



Pr(two lineages coalesce in the previous generation)

Example

$$\begin{aligned} \text{Pr(COAL in the previous generation)} &= \# \text{ of possible parents} \\ &\times \text{Pr(two lineages pick the same parent)} \end{aligned} \quad (9)$$

where $2N = 10$

Pr(two lineages coalesce in the previous generation)

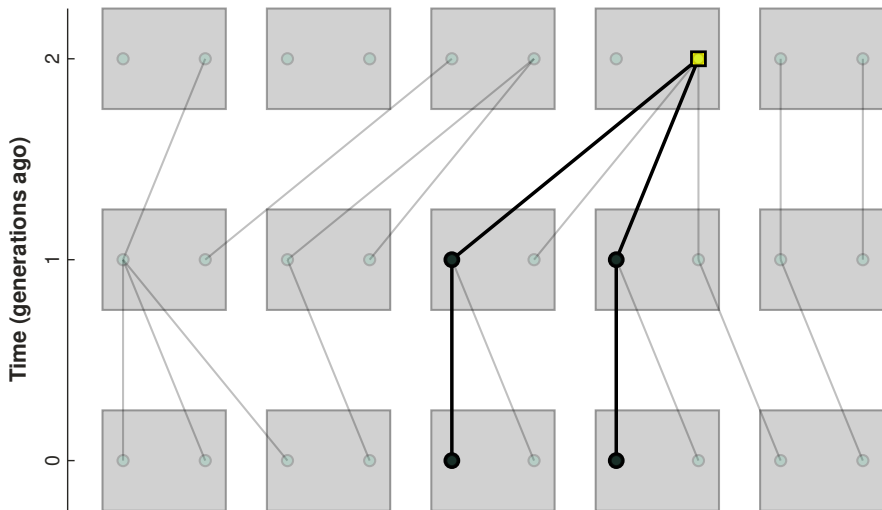
Example

$$\begin{aligned} \text{Pr}(\text{COAL in the previous generation}) &= \# \text{ of possible parents} \\ &\times \text{Pr}(\text{two lineages pick the same parent}) \end{aligned} \quad (9)$$

where $2N = 10$

$$\text{Pr}(\text{COAL in the previous generation}) = 10 \times \frac{1}{10} \times \frac{1}{10} = \frac{1}{2N} \quad (10)$$

Parent \neq Most Recent Common Ancestor (MRCA)



Pr(two lineages coalesce two generations ago)

Example

$$\begin{aligned} \Pr(\text{COAL two generations ago}) &= \Pr(\text{no COAL at generation one}) \\ &\quad \times \Pr(\text{COAL at generation two}) \end{aligned} \tag{11}$$

where $2N = 10$

Pr(two lineages coalesce two generations ago)

Example

$$\begin{aligned} Pr(\text{COAL two generations ago}) &= Pr(\text{no COAL at generation one}) \\ &\quad \times Pr(\text{COAL at generation two}) \end{aligned} \quad (11)$$

where $2N = 10$

$$Pr(\text{COAL two generations ago}) = \left(1 - \frac{1}{10}\right) \times \frac{1}{10} = \left(1 - \frac{1}{2N}\right) \times \frac{1}{2N} \quad (12)$$

Coalescent times for $k \rightarrow k - 1$ lineages

Defintion

$$Pr(\text{COAL at generation } t) = \left(1 - \frac{1}{2N}\right)^{t-1} \times \frac{1}{2N} \quad (13)$$

Given that $T_{k \rightarrow k-1} \sim \text{Geo}(p)$ where $p = \frac{1}{2N}$ the expected time to coalescence for $k \rightarrow k - 1$ lineages is:

Coalescent times for $k \rightarrow k - 1$ lineages

Defintion

$$Pr(\text{COAL at generation } t) = \left(1 - \frac{1}{2N}\right)^{t-1} \times \frac{1}{2N} \quad (13)$$

Given that $T_{k \rightarrow k-1} \sim \text{Geo}(p)$ where $p = \frac{1}{2N}$ the expected time to coalescence for $k \rightarrow k - 1$ lineages is:

$$\mathbb{E}(T_{k \rightarrow k-1}) = \frac{1}{p} = 2N \quad (14)$$

Overview

1. Quiz Review
2. Wright Fisher Model
3. Standard Coalescent Model
- 4. Derivations**