Topic 4: The Coalescent Process

Learning Objectives

1. What is the **Coalescent Process** and what does it describe?
   a. What does the i^{th} coalescent time represent and what is its distribution?
   b. What is the expected time to the i^{th} coalescent event?
   c. What is the variance in coalescent times?
   d. Draw an appropriately scaled coalescent genealogy
2. What is the relationship between coalescent times and population size?
3. Describe the assumptions of the infinite sites model of mutation
   a. What are three measures of genetic diversity in this model of mutation?
   b. Calculate the measures of genetic diversity from a given sample.
   c. What is the expected number of segregating sites in a sample?
   d. What is the expected number of pairwise differences in a sample?
4. Simulate a coalescent process for a sample of size n with and without mutation.
   a. Compare your simulations to the statistical expectations.

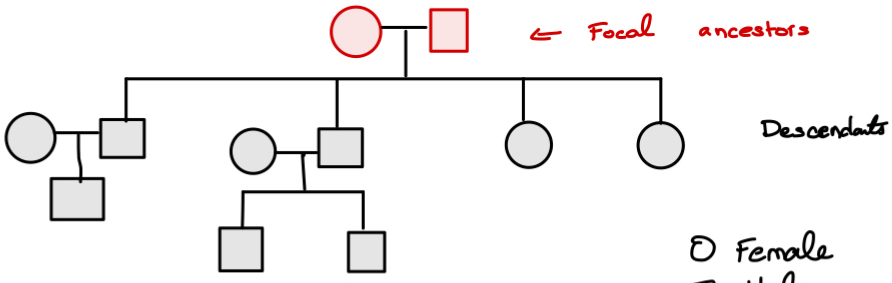
Lecture 4.1 The Kingman Coalescent

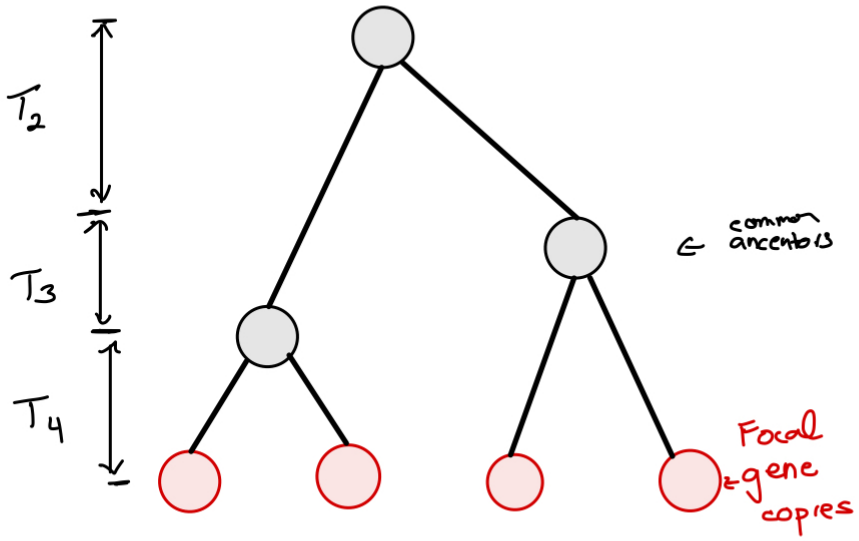
Genealogies

Genealogies are something that we are often familiar with conceptually from litterature, history, or DNA ancestry databases. However it is important to distinguish between a **pedigree** and a **genealogy**.

A pedigree is the graphical description of one or more ancestor(s) and its/their offspring. Nodes in a pedigree indicate *individuals* and edges represent parent-offspring relationships. In other words a pedigree is defined *forward-in-time* starting from historical ancestor and describing all subsequent descendants of that ancestor until the "present" day. The length of edges in a pedigree do not have meaning.

A genealogy is a graphical representation of a collection of *sampled descendants* and their common ancestors. Nodes in a genealogy represent *genetic samples* and their common ancestors. Edges represent lineages of inheritance from an ancestor to a descendent and the length of these edges are proportional to the time (measured in generations or scaled generations) between ancestors and descendants.





The **coalescent** process describes the genealogical history of a sample of genomes, specifically how samples coalesce into common ancestors. There are a few important things to keep in mind about a coalescent as we discuss it:

1. The coalescent considers a **sample** of n individuals from a large population of size N where $n \ll N$
2. The outcome of a coalescent process (like a poisson process) is on the **waiting times** (e.g., T_n, T_{n-1}, or T_2) between coalescent events.

Let T_i be the time over which there is exactly i distinct ancestors in the sample where $i = n, n-1, \ldots, 2$. Specifically we will show that if there are i lineages then the waiting time until the next coalescent event is exponenitally distributed with rate: $\lambda = \frac{\binom{i}{2}}{N}$

$$\Pr(T_i = t_i) = \frac{\binom{i}{2}}{N} e^{-\frac{\binom{i}{2}}{N} t_i}$$

3. The ultimate aim of the coalescent is to provide us with **statistical expectations** for how genomes in a sample should be related to one another.

The expected time to a coalescent event is:

$$E[T_i] = \frac{N}{\binom{i}{2}}$$

The variance in time to the coalescent event is:

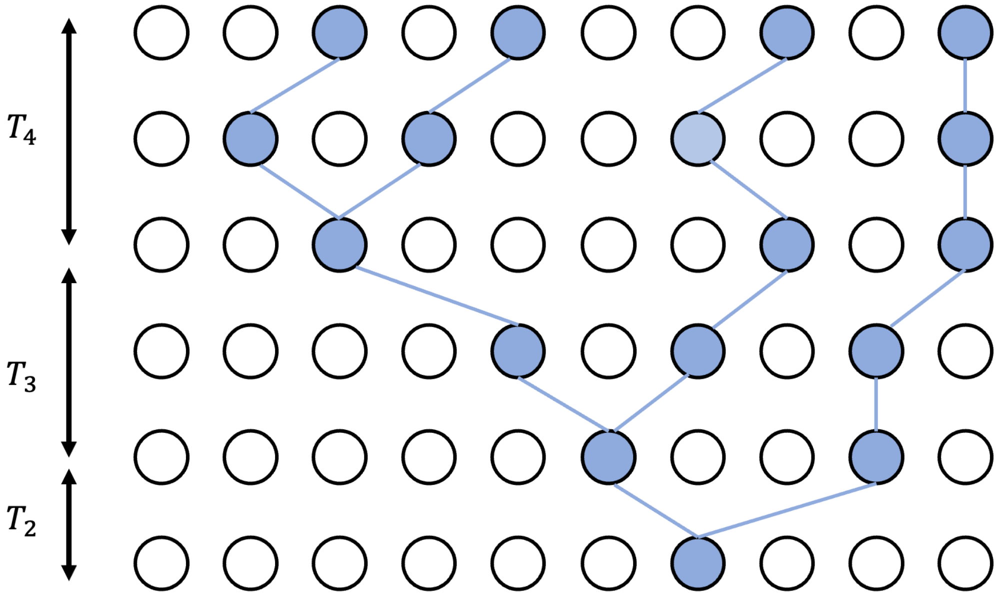
$$Var[T_i] = \frac{N^2}{\binom{i}{2}^2}$$

Derivation of the Coalescent

In this section we will derive the distribution of waiting times from the underlying biological process. To do so we use the Wright-Fisher model.

In Topic 2 we discussed the Wright-Fisher model as a process of simulating genetic drift in a population of size N going forward in time and how the number of individuals carrying different alleles 'A' and 'a' changed from generation to generation.

Here we are going to use the exact same process, but are going to work backward in time from offspring to parents. Rather than focus on different alleles ('A' and 'a') we are going to focus on what happens to a set of n gene copies that are in our sample versus the $N - n$ that are not.



The probability that i lineages have j ancestors in the immediately previous generation is:

$$G_{i,j} = \frac{S_i^{(j)} N_{\lfloor j \rfloor}}{N^i}$$

- $S_i^{(j)}$ is the "Sterling number of the second kind" and gives the number of ways a set of i elements can be partitioned into j subsets.
  - We have: $S_{i,i} = 1$ and $S_i^{(i-1)} = \binom{i}{2} = \frac{i(i-1)}{2}$
  - The value of $S_i^{(j)}$ for other values of i and j is calculated recursively
- $N_{\lfloor j \rfloor} = \underbrace{N(N-1)(N-2)\ldots(N-(j-1))}_{j \text{ terms}}$ is the "descending factorial"
- So we have: $G_{i,j} = {\color{red} S_i^{(j)}} {\color{blue} \frac{N_{\lfloor j \rfloor}}{N^i}}$
  - Blue: Probability a sample with i individuals has a specific set of j ancestors
  - Red: Number of ways to have j ancestors

The probability that i offspring have exactly i ancestors (No Coalescent Event) then is:

$$\begin{aligned} G_{i,i} =& \frac{N_{\lfloor i \rfloor}}{N^i} = \frac{\cancel{N}}{\cancel{N}} \frac{N-1}{N} \frac{N-2}{N} \ldots \frac{N-(i-1)}{N} \\ =& \left(1 - \frac{1}{N}\right)\left(1 - \frac{2}{N}\right) \ldots \left(1 - \frac{i-1}{N}\right) \\ =& 1 - \frac{\sum_{j=1}^{i-1} j}{N} + \mathcal{O}\left(\frac{1}{N^2}\right) \approx 1 - \frac{\binom{i}{2}}{N} \end{aligned}$$

The probability that i offspring have exactly $i-1$ ancestors in the previous generation (One Coalescent Event) then is:

$$
\begin{aligned}
G_{i,i-1} =& \frac{S_i^{(i-1)} N_{\lfloor i-1 \rfloor}}{N^i} = \frac{\binom{i}{2}}{N}\frac{N_{\lfloor i-1 \rfloor}}{N^{i-1}} \\
=& \frac{\binom{i}{2}}{N}\left[\left(1-\frac{1}{N}\right)\left(1-\frac{2}{N}\right)\ldots\left(1-\frac{i-2}{N}\right)\right] \\
=& \frac{\binom{i}{2}}{N}\left[1-\frac{\sum_{j=1}^{i-2} j}{N}+\mathcal{O}\left(\frac{1}{N^2}\right)\right] \\
=& \frac{\binom{i}{2}}{N}+\mathcal{O}\left(\frac{1}{N^2}\right) \approx \frac{\binom{i}{2}}{N}
\end{aligned}
$$

So if N is large there are only two options: either no coalescent event occurs in a given generation or exactly 1 coalescent event occurs. The probability of two or more coalescent events occurring simultaneously is $\mathcal{O}\left(\frac{1}{N^2}\right)$.

The probability of a coalescent event not happening in $[t]$ generations then is:

$$
G_{i,i}^t \xrightarrow{N\to\infty} e^{-\binom{i}{2}t}
$$

Lecture 4.2 Summary Statistics in the Coalescent

Scaling Time in the Coalescent

Recall from the previous lecture that the distribution of coalescent times is given by:

$$
\Pr(T_i = t_i) = \frac{\binom{i}{2}}{N} e^{-\frac{\binom{i}{2}}{N} t_i}
$$

Where T_i is the number of generations over which there are exactly i lineages in the sample.

Since N can be vary large and hence the coalescent rate $\binom{i}{2}/N$ very small the waiting time to the next coalescent event can be very long when measured in units of generations.

It is convenient to rescale time in the coalescent model into "coalescent units" by defining:

$$
\tau = \frac{t}{N} \quad t = N\tau
$$

τ is small relative to t

So we obtain:

$$
\Pr(\mathrm{T}_i = \tau_i) = \binom{i}{2} e^{-\binom{i}{2} t_i}
$$

Population Size Affects Coalescent Rate

When we talk about population size in the coalescent model this is not the **census population size** but rather the **effective population size**. Defining effective population size is beyond the scope of this course, but in general it is much much smaller than the census size.

Estimates of effective population sizes for humans vary widely and depended on geography. They can be estimated in a variety of ways each with different biases/strengths. Estimates for ancestory of European populations which are the most well studied genetically range from 10,000 (Takahata 1993) to <3,000 (Tenesa et al. 2007).

Discussion: Consider two populations, one with an $N_e = 100$ and one with $N_e = 200$. Suppose we sample $n = 10$ genomes from each population, in which population is a coalescent event expected to occur first?

Recall that the distribution of coalescent times is:

$$\Pr(T_i = t_i) = \frac{\binom{i}{2}}{N_e} e^{-\frac{\binom{i}{2}}{N_e} t_i}$$

an exponential distribution with mean:

$$E[T_i] = \frac{N}{\binom{i}{2}}.$$

The bigger the population the population size then the longer it takes a coalescent event to occur.

Similarly, as populations expand the coalescent rate slows down. As populations decline in size the faster the events occur.

Summary Statistics

0. Moments of the Distribution

This is an exponential distribution with rate $\binom{i}{2}$, so that we have the expected time the i^{th} coalescent event is:

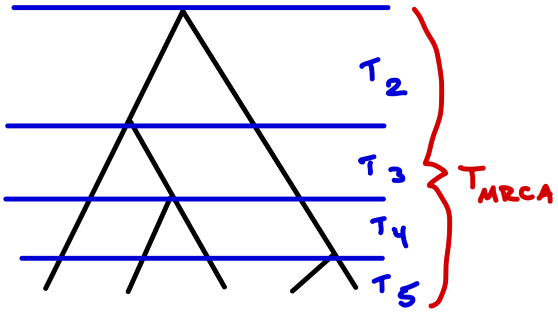
$$E[T_i] = \frac{1}{\binom{i}{2}} = \frac{2}{i(i-1)}$$

Similarly we have:

$$Var[T_i] = \frac{1}{\lambda^2} = \frac{1}{\binom{i}{2}^2}$$

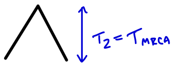
In addition to the time between each coalescent event we are also interested in three other quantities:

1. **The Time to the Most Recent Common Ancestor**

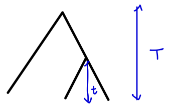


Example: T_{MRCA}

What is the T_{MRCA} when $n = 2$?



What is the T_{MRCA} when $n = 3$?

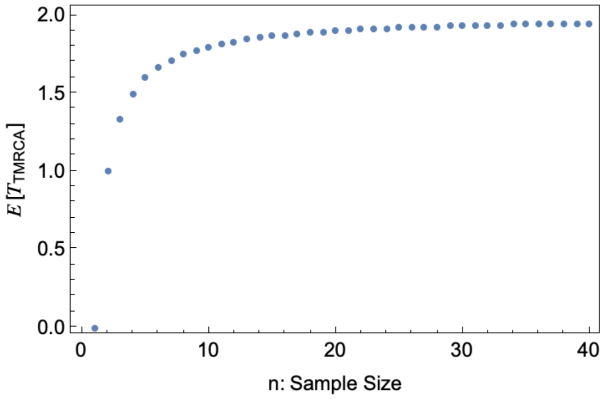


$T_{MRCA} = \int_0^T \lambda_3 e^{-\lambda_3 t} \lambda_2 e^{-\lambda_2 (T-t)} dt = \frac{\lambda_1 \lambda_2}{\lambda_1 - \lambda_2} e^{-T(\lambda_1 + \lambda_2)}$

$$T_{\text{MRCA}} = \sum_{i=2}^{n} T_i$$

The expected time to a common ancestor then is:

$$E[T_{\text{MRCA}}] = 2\left(1 - \frac{1}{n}\right)$$



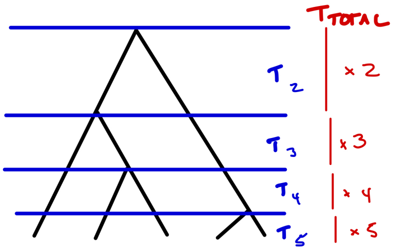
This asymptotes at 2.

$$\lim_{n\to\infty} 2\left(1 - \frac{1}{n}\right) = 2$$

The distribution of times to the common ancestor are:

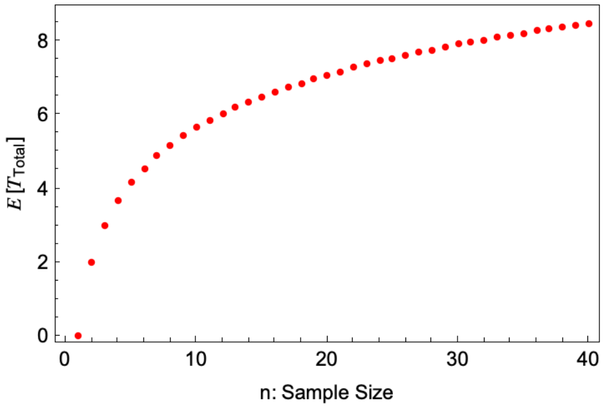
$$\Pr(T_{\text{MRCA}} = \tau) = \sum_{i=2}^{n} \binom{i}{2} e^{-\binom{i}{2}\tau} \prod_{j=2, j\neq i}^{n} \frac{\binom{j}{2}}{\binom{j}{2} - \binom{i}{2}}$$

2. **Total Branch Length**, T_{Total}



$$T_{\text{Total}} = \sum_{i=2}^{n} iT_i$$

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

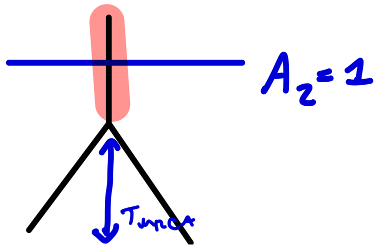


3. **The probability that there are k ancestors at time T.**

Building on the analogy between the exponential distribution in a Poisson process and time to the next coalescent event $\Pr(\mathrm{T}_i = \tau_i)$, we can also calculate the probability of having k ancestors (e.g., having $n-k$ coalescent events) at time T (analogous to the Poisson distribution in the Poisson process) and the time until the $n-k$th coalescent event (analgous to the Erlang distribution).

Let $A_n(t)$ be the number of ancestors at time τ.

- First consider when $k = 1$

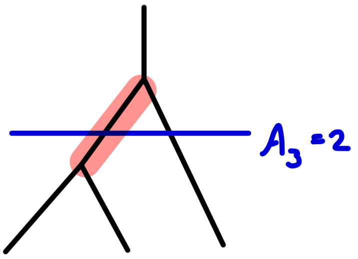


$$
\begin{aligned}
\Pr(A_n(\tau) = 1) &= \int_0^\tau \Pr(T_{\text{MRCA}} = x)dx \\
&= \int_0^\tau \sum_{i=2}^{n} \binom{i}{2} e^{-\binom{i}{2}x} \prod_{j=2, j\neq i}^{n} \frac{\binom{j}{2}}{\binom{j}{2} - \binom{i}{2}} dx \\
&= \sum_{i=2}^{n} \prod_{j=2, j\neq i}^{n} \frac{\binom{j}{2}}{\binom{j}{2} - \binom{i}{2}} \underbrace{\int_0^\tau \binom{i}{2} e^{-\binom{i}{2}x} dx}_{\text{cummulative dist.}} \\
&= \sum_{i=2}^{n} \prod_{j=2, j\neq i}^{n} \frac{\binom{j}{2}}{\binom{j}{2} - \binom{i}{2}} \left[1 - e^{-\binom{i}{2}\tau}\right]
\end{aligned}
$$

Because $\sum_{i=2}^{n} \prod_{j=2, j\neq i}^{n} \frac{\binom{j}{2}}{\binom{j}{2} - \binom{i}{2}} = 1$ this expression simplfies to:

$$
\Pr(A_n(\tau) = 1) = 1 - \sum_{i=2}^{n} e^{-\binom{i}{2}\tau} \prod_{j=2, j\neq i}^{n} \frac{\binom{j}{2}}{\binom{j}{2} - \binom{i}{2}}
$$

- Then consider when $k \geq 2$

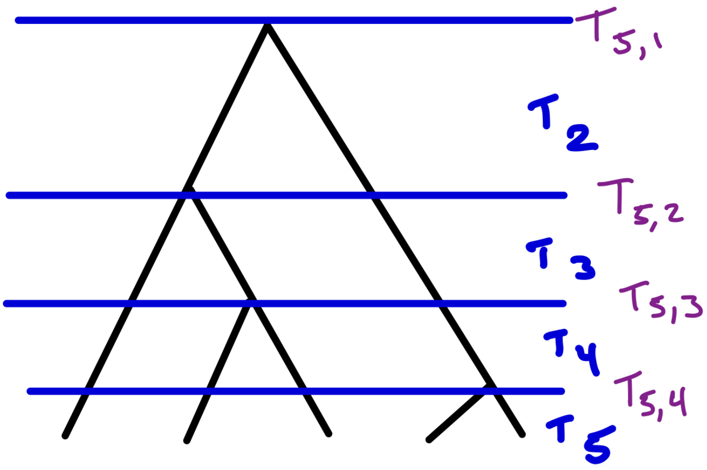


$$
\Pr(A_n(\tau) = k \geq 2) = \int_0^\tau \Pr(T_{n-k} = x) \int_{\tau - x}^{\infty} \Pr(T_{n-(k-1)} = y) dy dx
$$

With considerable algebra that is not enlightening we have:

$$
\Pr(A_n(\tau) = k \geq 2) = \frac{1}{\binom{k}{2}} \sum_{i=k}^{n} \binom{i}{2} e^{-\binom{i}{2}\tau} \prod_{j=k, j\neq i}^{n} \frac{\binom{j}{2}}{\binom{j}{2} - \binom{i}{2}}
$$

4. **Time to the $(n-k)^{\text{th}}$ coalescent event.**



Let $T_{n,k} = \sum_{i=k}^{n} T_i$, in other words the time until there are k lineages. We obtain an expression similar to that for the time to the most recent common ancestor:

$$\Pr(T_{n,k} = \tau) = \sum_{i=k+1}^{n} \binom{i}{2} e^{-\binom{i}{2}t} \prod_{j=k+1, j\neq i}^{n} \frac{\binom{j}{2}}{\binom{j}{2} - \binom{i}{2}}$$

The Shape of a Coalescent

In this section we are trying to gain an intuition for the expected shape of coalescent genealogy so that we can "draw them to scale". Remember that the coalescent in a random process so any one realization of this process looks different, but we can study the expected shape.

Example: How long does the first coalescent event take?

Consider a sample of size n, the expected time to the most recent common ancestor (in coalescent units) is:

$$E[T_{\text{MRCA}}] = 2\left(1 - \frac{i}{n}\right)$$

1. How long does the first coalescent event take relative to the time tot he MRCA?

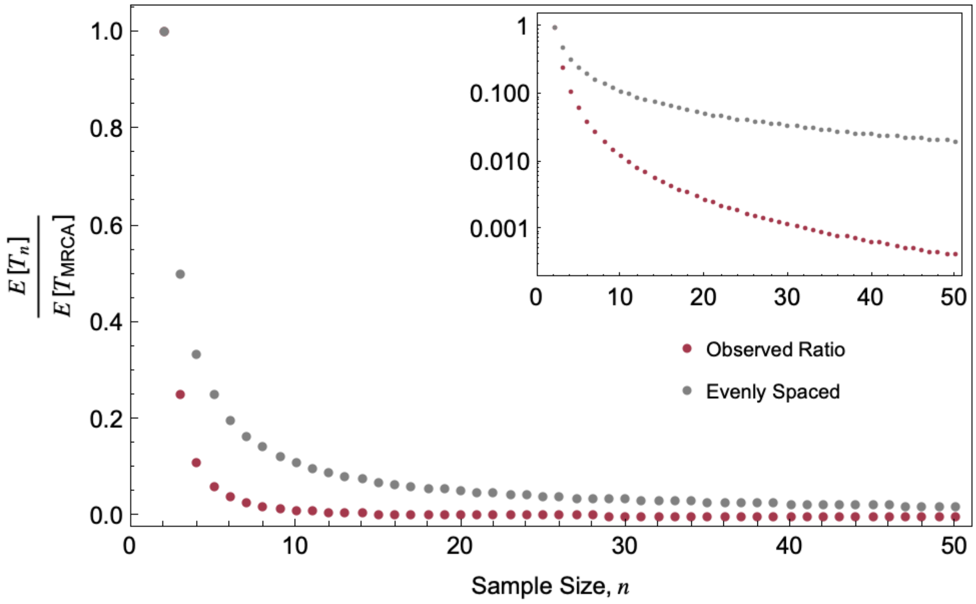
We have that the time to the first coalescent event is:

$$E[T_n] = \frac{2}{n(n-1)}$$

So

$$\frac{E[T_n]}{E[T_{MRCA}]} = \frac{\frac{2}{n(n-1)}}{2\left(1 - \frac{1}{n}\right)} = \frac{1}{(n-1)^2}$$

If the coalescent events were evenly spaced we would expect the each event to take $\frac{1}{n-1}$ of the time.



The key takeaway here is that for even moderate n the first coalescent event takes up only a very small portion of the time and much less then we would expect if the events were evenly spaced.

Example: How long does the last coalescent event take?

Now let's ask the same question but in terms of the last coalescent event, T_2.

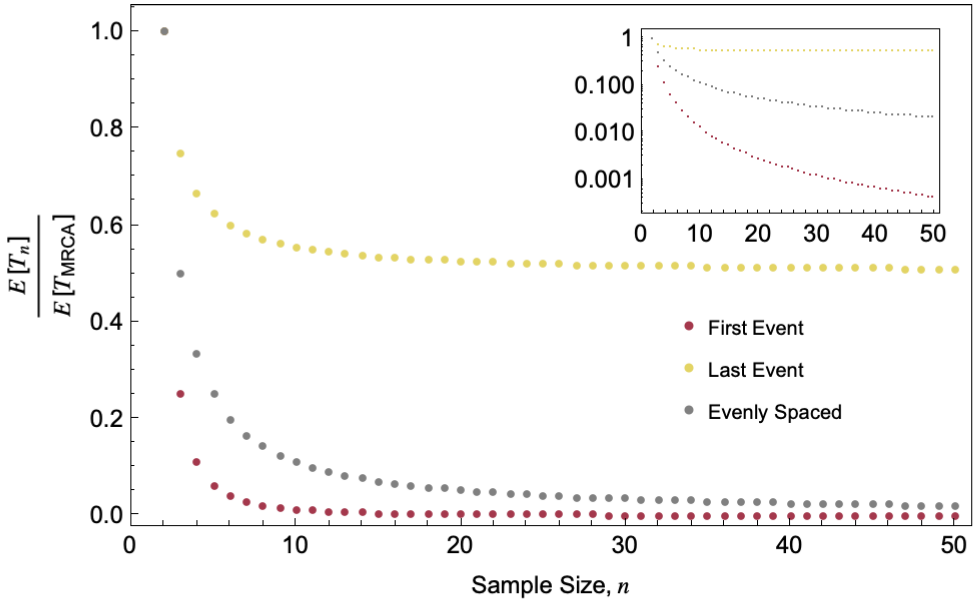
1. How long does the last coalescent event take?

$$E[T_2] = \frac{2}{2(2-1)} = 1$$

This is another way of defining the coalescent time unit, the expected time until the coalescence of two samples!

2. How long does the last coalescent event take relative to the MRCA?

$$\frac{E[T_2]}{E[T_{MRCA}]} = \frac{1}{2\left(1-\frac{1}{n}\right)} = \frac{n}{2(n-1)}$$



In the limit as $n \to \infty$,

$$\frac{E[T_2]}{E[T_{MRCA}]} = \frac{n}{2(n-1)} = \frac{1}{2}$$

So the last event takes up *atleast* half of the time!

Discussion: Use these results to sketch an appropriately scaled coalescent

We can use this last result to intuitively sketch a coalescent. Given the TMRCA, approximately, half of the time is spent on the last event, 1/4 of the time of the second to last event, 1/8th on the third, etc.

Lecture 4.3 Human Neanderthal Couples

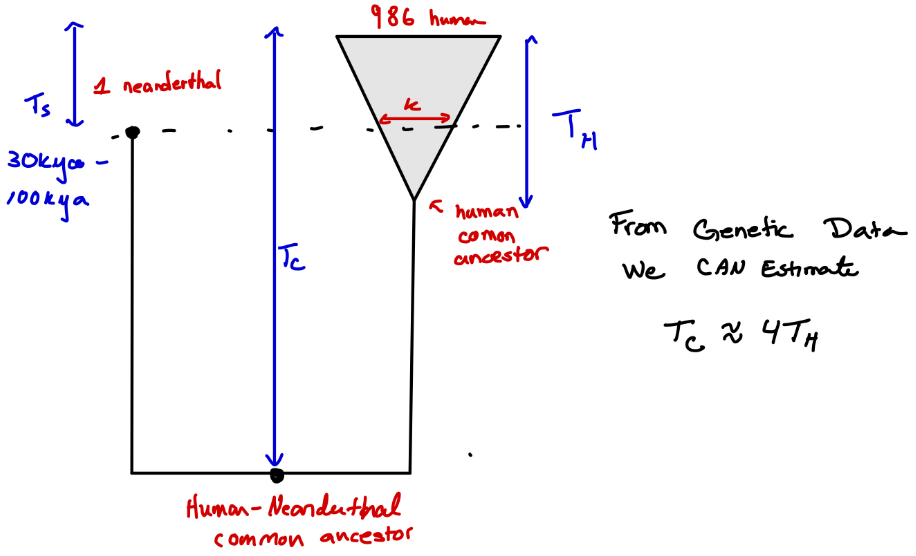
Analyzing Human-Neanderthal Genetic Data

Neanderthals are an extinct hominid group that is known to have coexisted with humans as recently as 30kya.

Did ancient humans and Neanderthals interbreed randomly?

Discussion: What does it mean to interbreed randomly in the Wright-Fisher Model?

Nordberg (1998) addressed this question using the first Neanderthal mitocondrial sequence and 986 human mtDNA sequences available at that time. We can sketch out a hypothetical genealogy of these 987 sequences as:



- We define the time to the common ancestor of the 986 human samples as T_H
- The focal Neanderthal lived T_S time ago. Note that T_S may be greater than or smaller than T_H
- The time to the common ancestor of the humans *in our sample* and the Neanderthal *in our sample* is T_C

Genetic data (we will discuss this more in the next lecture) suggests that $T_C \approx 4T_H$.

Question: What is the $\Pr(T_C \geq 4T_H | T_S)$ if there had been random mating between humans and Neanderthals?

Step 1: First we need to estimate T_S. Time here is measured in "Coalescent Generations".

- Suppose that the focal Neanderthal lived between 30kya
- The human generation time is approximately 20years.
- The "effective" population size of humans is approximately $N = 3,400$. Effective population size is a complex quantity but here it represents the appropriate average of historical population sizes over time and the fact that mitocondrial DNA is passed on by only females.
  - As a result we can estimate:

$$T_S \approx \frac{30,000}{3400 * 20} = 0.44$$

Step 2: Express the probability in terms of the number of human ancestors at time T_S.

- We do not know T_H it may be bigger or smaller than T_S. Specifically, we have to consider the possiblity that there are k ancestors of our human sample present at time T_S.
    - If $k=1$ then $T_H < T_S$
    - If $k>1$ then $T_H > T_S$
        - in this case we only want to consider "trees" where the k human ancestors coalesce before the coalesce with the Neanderthal sample.
- Let $A_n(\tau)$ be the number of ancestors of a sample of size n at time τ in the past.

$$\Pr(T_C \geq 4T_H|T_S) = \sum_{k=1}^{968} \Pr(A_{968}(T_S)=k)\,\Pr(\text{tree}|k)\,\Pr(T_C \geq 4T_H|k,\text{tree})$$

Step 3: Calculating the probability that there are k ancestors, $Pr(A_{968}(T_S)=k)$

$$\begin{aligned}\Pr(A_{968}(T_S)=1) =& 1 - \sum_{i=2}^{968} e^{-\binom{i}{2}T_S} \prod_{j=2,j\neq i}^{n} \frac{\binom{j}{2}}{\binom{j}{2}-\binom{i}{2}} \\ \Pr(A_{968}(T_S)=k\geq 2) =& \frac{1}{\binom{k}{2}} \sum_{i=k}^{968} \binom{i}{2} e^{-\binom{i}{2}T_S} \prod_{j=k,j\neq i}^{968} \frac{\binom{j}{2}}{\binom{j}{2}-\binom{i}{2}}\end{aligned}$$

- For $T_S = 0.44$ we have: $E[A_{968}(T_S)] = 4.68$ so there was likely 4 to 5 human lienages present at time T_S.

Step 4: Calculate the probability that the humans share a common ancestor before they coalesce with the Neanderthal, $\Pr(\text{tree}|k)$

- We are interested in the "chance that the entire set of human samples shares a common ancestor to the exclusion of the Neanderthal"
- In other words, at time T_S we now have a sample of $k+1$ lineages in which one sample is "labeled"
    - If there are j lineages one of which is labeled there are $\binom{j}{2}$ possible paris that can coalesce in the next event. $j-1$ of thse events will include the labeled lineage.
    - The probability that the next coalescent event doesn't involve the labeled lineage then is: $\left(1-\frac{j-1}{\binom{j}{2}}\right)$
- The probability we want then is just the product for $j=k+1$ down to, and including, $j=2+1=3$

$$\Pr(\text{tree}) = \prod_{j=3}^{k+1}\left(1-\frac{j-1}{\binom{j}{2}}\right) = \frac{2}{k(k+1)}$$

- We have $E[\Pr(\text{tree})] = 0.085$, so 8% of the time we obtain the tree we want.

Step 5: Calculate the relative values of T_H and T_C, $\Pr(T_C \geq 4T_H|k,\text{tree})$

$$\Pr(T_C \geq 4T_H|T_S,\text{tree}) = \Pr(T_C - 4T_H \geq 0|\text{tree},k)$$

- Note that k and tree are calculated relative to T_S but T_C and T_H relative to the present day so let's adjust so everything is in terms of T_S.

$$\Pr(T_C - 4T_H \geq 0|k,\text{tree}) = \Pr((T_C - T_S) - 4\,(T_H - T_S) \geq 3T_S|k,\text{tree})$$

Here $(T_C - T_S) = \sum_{i=2}^{k+1} T_i$ is the time until $k+1$ lineages coalesce and $(T_H - T_S) = \sum_{i=3}^{k+1} T_i$ is the time over which the first k of these lineages coalesce.

$$\begin{aligned}
&\Pr\left(\left(T_C - T_S\right) - 4\left(T_H - T_S\right) \geq 3T_S | k, \text{tree}\right) \\
&= \Pr\left(\sum_{i=3}^{k+1} T_i - 4\sum_{i=3}^{k+1} T_i \geq 3T_S\right) \\
&= \Pr\left(T_2 - 3\sum_{i=3}^{k+1} T_i \geq 3T_S\right) \\
&= \Pr\left(\frac{T_2}{3} - \sum_{i=3}^{k+1} T_i \geq T_S\right)
\end{aligned}$$

This quantity can be calculated numerically given the distribution of coalescent times.

Solution:
Putting all the pieces together we have $\Pr(T_C \geq 4T_H) = 0.0063$. We would only have a 0.6% chance of observing our data if humans and Neanderthals mated randomly.

Lecture 4.4 The Coalescent with Mutation

The independence of coalescence and mutation

Recall that the Wright-Fisher model is a model of **neutral genetic drift**. This means that the Kingman coalescent is a model of neutral drift. Hence any mutations that occur between parent and offspring are *neutral* and have no impact on who has offspring/how parents are chosen.

The fact that mutations are assumed to be neutral is very useful. This means that given a coalescent genealogy, we can randomly simulate mutations on top of this genealogy to obtain expressions for the genetic diversity of a sample of "genomes". This is where the real power of the coalescent comes in:

The coalescent provides us with statistical expectations for the genetic diversity of a sample under a given demographic (e.g., N) model.

Specifically, we want to combine the coalescent stochastic process described in the past few lectures with a second stochastic model for the occurrence of mutations. There are two common models of mutation used in coalescent theory each of which makes different simplifying assumptions:

1. The **infinite sites model**: assumes that every new mutation that occurs occurs at a different "site" in the genome. In other words we can model the ancestor of a sample with a long string of 0's, each new mutation that occurs changes a different one of these 0's to a 1.

This model is appropriate for modelling mutations at neucleotides in large genomes.

2. The **infinite alleles model**: assumes that every new mutation that occurs creates a new "allele" that is distinct from all others.

Infinite Sites Model

Recall that time in the coalescent is measured in units of N generations, where N is the "effective population size". This means that the rate at which mutations occur along the branches must also be measured in these units.

Specifically let μ be the rate at which mutations occur at a given site in the genome per generation. In humans, $\mu \approx 1 \times 10^{-9} \frac{\text{mut}}{\text{site} \times \text{gen}}$.

While humans don't have an infinite number of sites in their genome, a medium sized gene in the human population is in the realm of 30,000 base pairs (e.g., sites) long.

The effective human population size varies depending on the region you are in, but is remarkably small, approximately 5,000 or on the order of 10^3.

So mutations in a meidum size gene accumulate at approximately a rate of:

Finally, for historical reasons the mutation rate in the coalescent models, θ, is multiplied by 2.

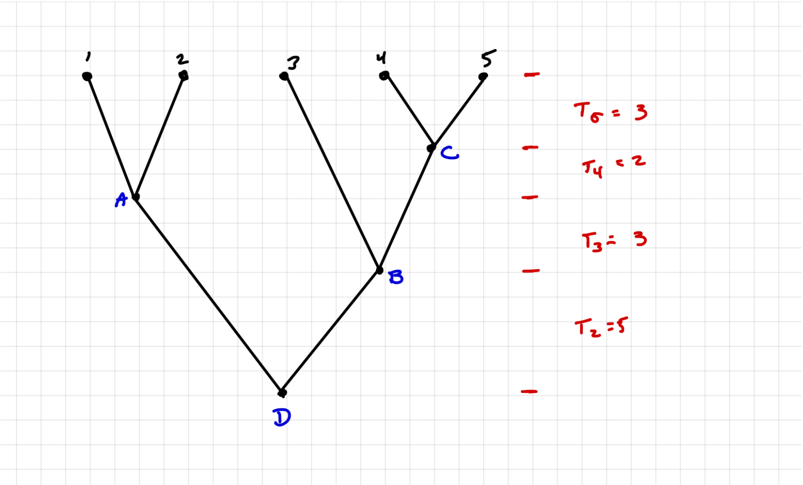
$$\theta \approx 1 \times 10^{-9} \frac{\text{mut}}{\text{site} \times \text{gen}} \times 3 * 10^4 \text{sites} \times 5 * 10^3 \frac{\text{gen}}{\text{coal}} \times 2 = 30 \times 10^{-2}$$

This is a very hand wavy estimate but θ is around 1.

Given this coalescent mutation rate, the key element to modelling mutation in the infinite sites model is noting that mutations occur as a Poisson process along the "edges" in a genealogy.

Example: Simulating the Infinite Sites model

Consider 5 sampled sequences related according to the genealogy shown. The ancestors are labled A-D.



1. If the mutation rate is $\theta = 1$ what is the expected number of mutations along each edge in the genealogy?

The number of mutations that occur along an edge of length τ should be Poisson distributed with rate $\lambda = \frac{\tau\theta}{2}$ (where we have to divide by 2 becuase θ is twice the mutation rate). The mean of this Poisson distribution is $\lambda = \frac{\tau\theta}{2}$

Edge	Length of Edge	Expected # of Mutations
1-A	5	2.5
2-A	5	2.5
3-B	8	4
4-C	3	1.5
5-C	3	1.5
C-B	5	2.5
A-D	8	4
B-D	5	2.5

2. Consider sequences 1 and 2 who share a common ancestor $\tau = 5$ coalescent units of time ago. What is the expected number of sites that differ between sample 1 and sample 2? What is the full distribution of the number of differences between these two samples?

Differences between sequences 1 and 2 include all mutations that occur between 1 and A and those that occur between A and 2. In other words samples 1 and 2 are seperated by a total of 10 coalescent units.

The number of mutations that occurs between the two samples then is $n \sim \mathcal{P}oi(10 * \theta/2)$ and the expected number of differences between sequence 1 and sequence 2 is 5.

There are three different measures of genetic diversity that are develope for the infinite sites model.

(Average) Number of Pairwise Differences, π

The first measure of genetic diversity we have is the number of pairwise differences $\pi_{i,j}$ between sampled sequences i and j.

Rather than follow the number of differences between each pair of samples, matheamtically we focus on a summary of these values by calcualting **average number of pairwise differences**, π

$$\pi = \frac{1}{\binom{n}{2}} \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} \pi_{i,j}$$

What is the **expected value** of π?

$$E[\pi] = E\left[\frac{1}{\binom{n}{2}} \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} \pi_{i,j}\right] = \frac{1}{\binom{n}{2}} \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} E\left[\pi_{i,j}\right]$$

Let $T_{i,j}$ be the coalescent time of sequences i and j. Since mutations may accumulate between i and the common ancestor and j and the common ancestor.

$$E[\pi] = \frac{1}{\binom{n}{2}} \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} E\left[\frac{\theta}{2} 2T_{i,j}\right] = \frac{\theta}{\binom{n}{2}} \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} E\left[T_{i,j}\right]$$

This leaves us with needing to calculate the expected pairwise coalescent time $E\left[T_{i,j}\right]$. Rather than derive this quantity from scratch, we will intuit its answer.

Suppose we have only two samples $n = 2$, the expected coalescent time between them is $\binom{n}{2} = 1$. How long it takes a pair of samples to coalesce should be independent of when and if they coalesce with other lineages. Hence intuitively $E\left[T_{i,j}\right] = 1$. Hence:

$$E[\pi] = \frac{\theta}{\binom{n}{2}} \times \binom{n}{2} = \theta$$

The Number of Segregating Sites, S

The second metric of genetic diversity in this model is the "number of segregating sites", S, which is the number of sites in the sampled sequence at which there has been a mutation.

Given that each mutation is assumed to occur at a different site in the infinite sites model, the number of segregating sites in this model is equal to the number of mutation that have occurred.

The number of mutations that occurs is a Poisson distributed random variable with rate $\lambda = \frac{\theta T_{Total}}{2}$, where T_{Total} is the total branch length in the genealogy. From lecture 4.2 we have:

$$T_{\text{Total}} = \sum_{i=2}^{n} iT_i$$

and

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

Hence the expected number of segregating sites is:

$$E[S] = 2\left(\sum_{i=1}^{n-1}\frac{1}{i}\right)\frac{\theta}{2} = \theta\sum_{i=1}^{n-1}\frac{1}{i}$$

Example: Simulating the Infinite Sites model cont.

1. What is the observed number of segregating sites in the following example sequences?

Seq	1	2	3	4	5	6	7	8	9	10
1	A	T	C	G	T	A	C	G	A	T
2	A	T	C	G	T	A	C	G	A	T
3	A	C	C	G	G	A	C	C	A	A
4	A	T	C	G	G	A	C	C	A	T
5	A	C	C	G	G	A	C	C	A	A

The sequences have four segregating sites: Bases 2,5,8, and 10

2. What is the expected number of segregating sites in a genealogy with $n = 5$ samples?

$$E[S] = \theta\sum_{i=1}^{5-1}\frac{1}{i} = \theta \times 2.08$$

Site Frequency Spectra η_i

- There is a third measure of genetic diversity called the site-frequency spectrum which is a discrete distribution given by $\xi_i \quad i \in \{1, 2, \ldots n\}$ which is the number of sites in the genome where the "mutant" allele is present. Note the domain:

Discussion: Why is the domain of ξ between 1 and $n-1$?

Example: Simulating the Infinite Sites model cont.

1. What is the observed site frequency spectrum in the data given above?

Lecture 4.5 Simulating the Coalescent

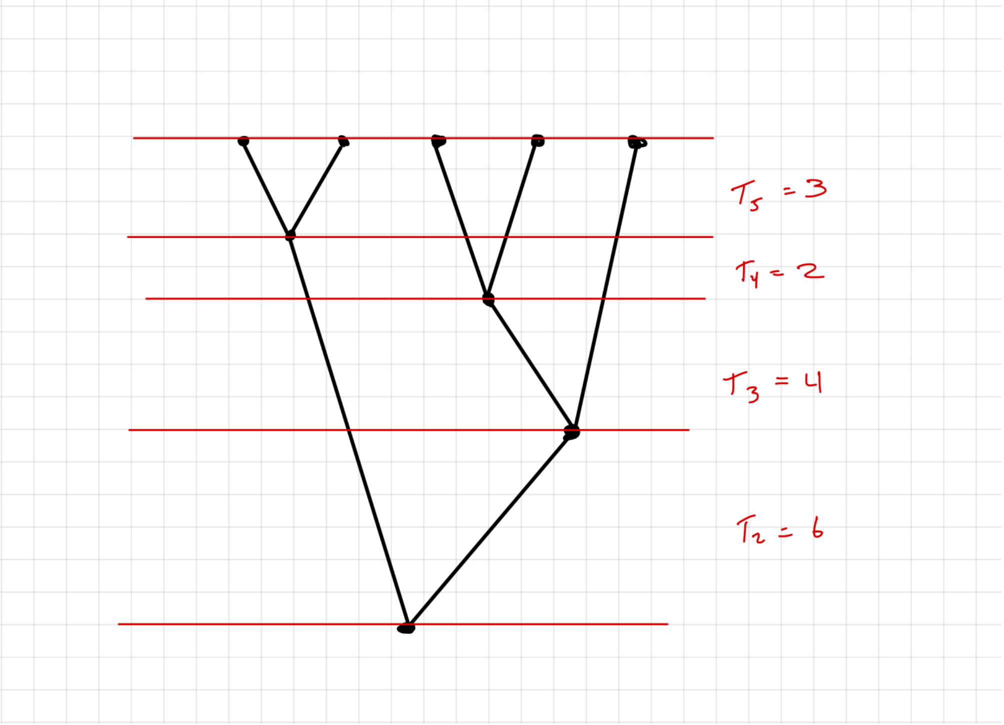
Python: Lecture4_5.ipynb

There are three key elements to simulating a coalescent:

- Note that the we can first simulate coalescent times and then the **topology** of the coalescent. Topology describes *who coalesced with whom*.
- Coalescent times are drawn from the **Kingman coalescent distribution**
- The **topology** of the coalescent can be stored by first labeling the lineages from 1 to n and then constructing the **cophenetic matrix**. The easiest way to describe this matrix is that $M_{i,j}$ describes the amount of shared ancestry between lineages i and j.

Example: Chphenetic matrices

1. What is the cophenetic matrix for the following genealogy?



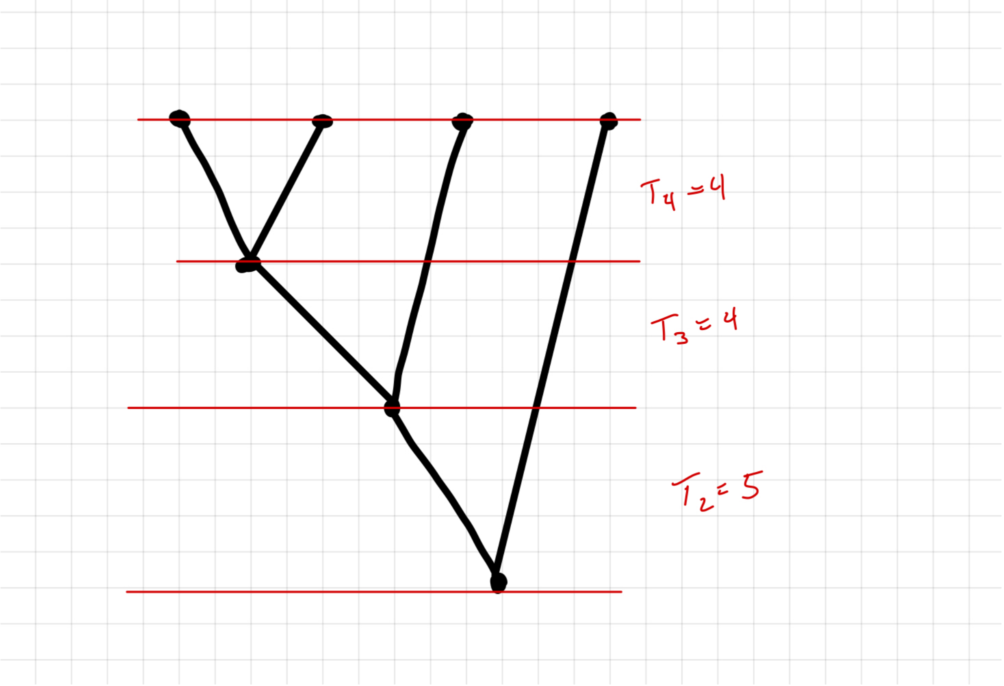
Solution:

$$M = \begin{bmatrix} 15 & 12 & 0 & 0 & 0 \\ 12 & 15 & 0 & 0 & 0 \\ 0 & 0 & 15 & 10 & 6 \\ 0 & 0 & 10 & 15 & 6 \\ 0 & 0 & 6 & 6 & 15 \end{bmatrix}$$

2. Draw the genealogy represented by the following Cophenentic matrix.

$$M = \begin{bmatrix} 13 & 9 & 5 & 0 \\ 9 & 13 & 5 & 0 \\ 5 & 5 & 13 & 0 \\ 0 & 0 & 0 & 13 \end{bmatrix}$$

Solution:



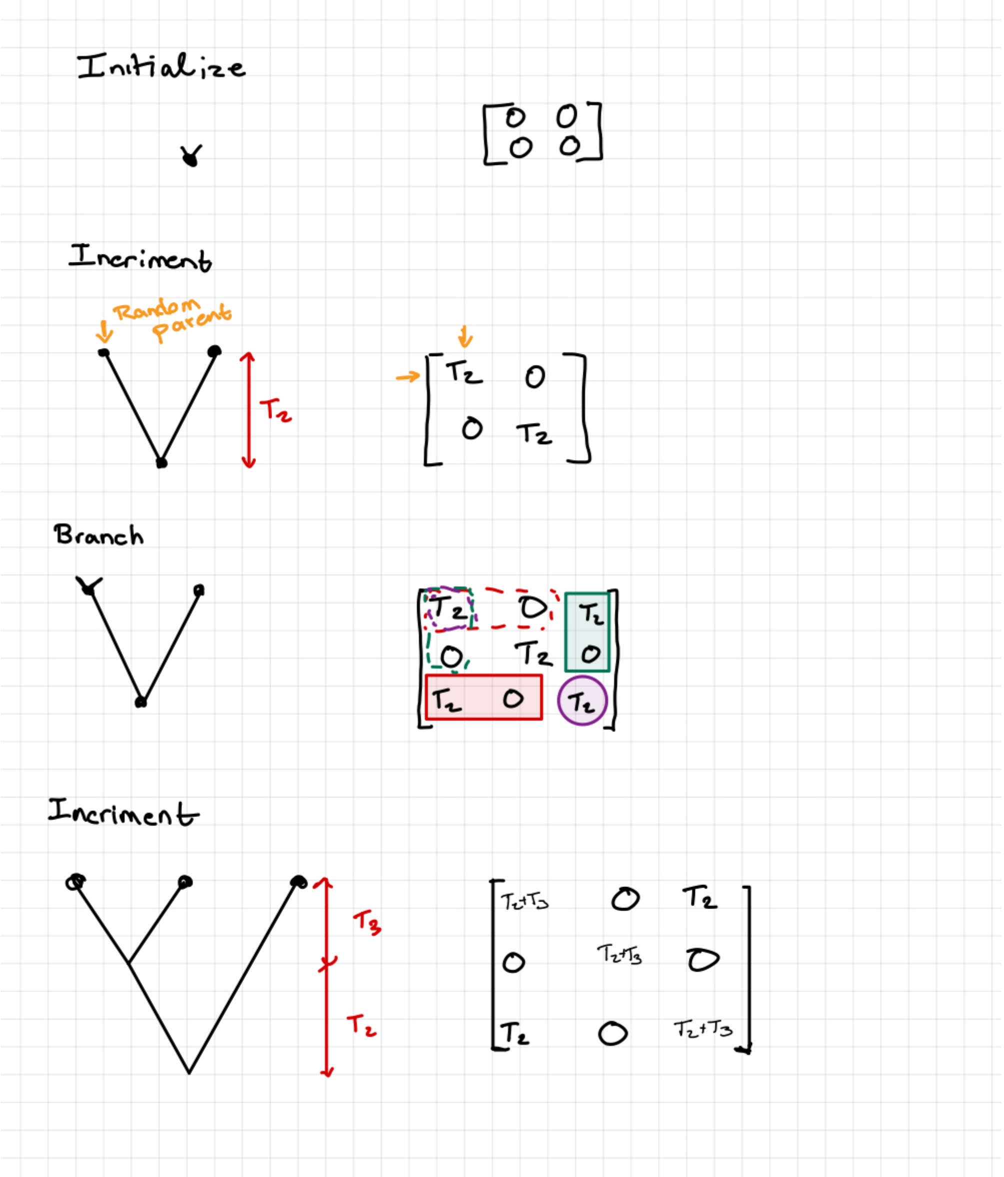
Simulating a Cophenetic Matrix

Given a sequence of coalescent times $[T_2, T_3, \dots, T_n]$ we can simualte the topology of a coalescent by creating a corresposnding cophenetic matrix **forward-in-time**.

- Initialization: At the base of the geneaology we have the following initial cophentic matrix:

$$M(0) = \begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix}$$

- For each coalescent time $T_i \in \{T_2 \dots T_n\}$
    - Increment time: Add T_i to each diagonal element
    - Add lineage:
        - Choose a parent lienage that branches at random from the current matrix of size $m \times m$ (say this is lineage $j \in \{1, 2, \dots, m\}$)
        - If $m < n$: Add a new daughter lineage by adding a $m + 1$ row and column:
            - Copying row j and column j to the matrix
            - Set $T_{m+1,m+1} = T_{j,j}$



Why it all Matters

We have spent several lectures talking about the coalescent. So let's take some time to assess what the coalescent is useful for. Genome sequencing is a powerful tool for understanding the demographic and evolutionary history of a population. The coalescent is one model that helps us interpret genomes by providing us with statistical expectations of how genetically diverse (for example by the site-frequency spectra) our sample should be. This is by far not the only way to use genomes to understand evolutionary history but it is a powerful one!

Coalescent theory has been behind many recent scientific discoveries about human ancestry, historical movement, and even the ancestral origins of disease.
https://www.nytimes.com/2024/01/10/science/ancient-human-genes-multiple-sclerosis.html