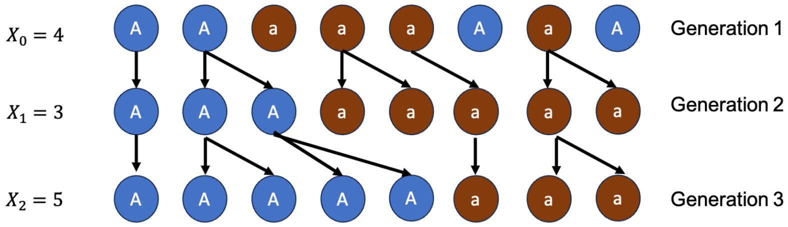
Topic 5c: Continuous Time/Discrete State

Lecture 5.7 Diffusion Approximations

Review: WF Model

Recall from topic 2 that the WF model is a DTDSMC describing neutral genetic drift in a small population. We can depict the dynamics of the WF model using the following schematic:



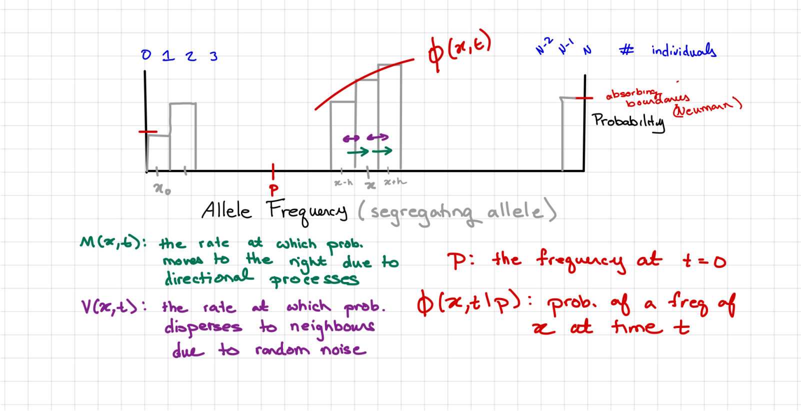
The probability of transitioning between having i 'A' parents to j 'A' offspring is:

$$\begin{aligned} P(X_{t+1} = j | X_t = i) = p_{ij} &= \binom{N}{j} \frac{i}{N}^{j} \left(\frac{N-i}{N}\right)^{N-j} \\ &= \underbrace{\binom{N}{j} p^j \left(1-p\right)^{N-j}}_{\mathcal{B}_j(N,p)} \end{aligned}$$

When the total population size N is reasonably small we can easily simulate this process, evaluate the dynamics numerically, and solve for the time to absorption/probability of absorption. But as N gets bigger these things become harder to do as the transition probability matrix gets too big.

The diffusion approximation is an approximation of the WF model for large (but still finite!) populations. The result of the diffusion approximation is a PDE of the function $\phi(x, t|p_0)$ describing the probability that a population has a (segregating) allele frequency x at time t given that the population stated with an allele frequency of p_0. This ODE is so widely used that it has its own name the "Kolmogorov Forward Equation".

Deriving the Kolmogorov Forward Equation


0 1 2 3
$\phi(x,t)$
N-2 N-1 N
# individuals
absorbing boundaries (Neumann)
Probability
x_0
p
x-h
x
x+h
Allele Frequency (segregating allele)
M(x,t): the rate at which prob. moves to the right due to directional processes
V(x,t): the rate at which prob. disperses to neighbours due to random noise
p: the frequency at t=0
$\phi(x,t|p)$: prob. of a freq of x at time t

To derive the diffusion approximation to the Wright Fisher model we begin by assuming that the number of individuals is large such that the probability that there are $0 < i < N$ individuals in the population with the 'A' allele can be approximated as a continuous frequency $0 < x < 1$. Note that space is subdivided into rectangles of width $\frac{1}{N}$ which are centred offset from the counts of individuals. We do not then consider the extreme cases of $i = 0$ and a $i = N$ and hence this approximation gives us the distribution of segregating allele frequencies.

Let $\phi(x, t|p)$ then be the probability that a 'A' allele that is segregating in the population at time has frequency x at time t given that the frequency starts at p at time 0. Probability can move either due to directional processes (e.g., selection, migration,

mutation) or **diffuse** in an arbitrary direction due to random drift.
Let's define:

- $m(x,t)$ as the probability that the process moves to the right due to directional processes
- $v(x,t)$ the probability that the process moves to the left ($\frac{1}{2}$ of the time) or to the right ($\frac{1}{2}$ of the time) due to random processes.

We can write the change in the volume of probability as:

$$\begin{aligned}\frac{\phi(x,t+\Delta t|p)\Delta x-\phi(x,t|p)\Delta x}{\Delta t} =&\phi(x,t|p)\Delta x-\phi(x,t|p)\Delta x\\ &-\left(m(x,t)\frac{\Delta t}{\Delta t}+v(x,t)\frac{\Delta t}{\Delta t}\right)\phi(x,t|p)\Delta x\\ &+\frac{1}{2}v(x-\Delta x)\frac{\Delta t}{\Delta t}\phi(x-\Delta x,t|p)\Delta x\\ &+\frac{1}{2}v(x+\Delta x)\frac{\Delta t}{\Delta t}\phi(x+\Delta x,t|p)\Delta x\\ &+m(x-\Delta x)\frac{\Delta t}{\Delta t}\phi(x-\Delta x,t|p)\Delta x\end{aligned}$$

Let

- $M(x,t)=E[\Delta x]$ be the expected amount of movement to the right

$$M(x,t)=m(x,t)*\Delta x$$

- $V(x,t)=\mathrm{Var}\left[\Delta x\right]=E[\Delta x^2]-E[\Delta x]^2$ be the variance in the movement due to random processes.

$$V(x,t)=\left(\frac{v}{2}\Delta x^2+\frac{v}{2}(-\Delta x)^2\right)-\underbrace{\left(\frac{v}{2}\Delta x+\frac{v}{2}(-\Delta x)\right)^2}_{=0}=\Delta x^2 v(x,t)$$

$$\begin{aligned}\frac{d\phi(x,t)}{dt}\frac{\Delta x}{\Delta x}=&-\left(\frac{\frac{M(x,t)}{\Delta x}\phi(x,t|p)\Delta x-\frac{M(x-\Delta x,t)}{\Delta x}\phi(x-\Delta x,t|p)\Delta x}{\Delta x}\right)\\ &+\frac{1}{2}\frac{\frac{V(x+\Delta x,t)}{\Delta x^2}\phi(x+\Delta x,t|p)\Delta x-\frac{V(x,t)}{\Delta x^2}\phi(x,t|p)\Delta x}{\Delta x}\\ &+\frac{1}{2}\frac{\frac{V(x-\Delta x,t)}{\Delta x^2}\phi(x+\Delta x,t|p)\Delta x-\frac{V(x,t)}{\Delta x^2}\phi(x,t|p)\Delta x}{\Delta x}\end{aligned}$$

The first row is the can be simplifed using the definition of the one-step derivative. The second two rows can be simplified using the definition of the two-step derivative.

$$\frac{d\phi(x,t)}{dt}=-\frac{\partial M(x,t)}{\partial x}+\quad\frac{1}{2}\frac{\partial V(x,t)}{\partial x}$$

Where the first term captures the change in the allele frequency due to directional processes (this is known as *mathematical drift*) and $M(x,t)$ is the infinitesimal mean.

The second term captures the change in the allele frequency due to random processes (this is known as *mathematical diffusion* or *biological drift*) and $V(x,t)$ is the **infinitesimal variance**.

The Wright-Fisher Model in Review

- The Wright-Fisher Model is a DTDS model of neutral genetic drift capturing the change in allele frequencies from generation to generation due to the random sampling (of parents).
- The WF model can be used to follow the number/frequency of a 'A' allele in a finite population.
- In the absence of mutation the WF model has two absorbing states: fixation of the 'A' and loss of the 'a'. We can use first-step analysis to calculate the relative probabilities of ending of in each of these absorbing states.

- In the presence of mutation, we can derive the stationary distribution of the WF model capturing the amount of time the population stays in a particular state in the long term.
- The Moran model is an alternative to the WF model capturing evolution in approximately continuous time. There are notable differences (by a factor of 2) in the rate of genetic drift between the WF and Moran models. Whether we model in discrete or continuous time matters!!
- The Kingman coalescent uses the WF model to describe the genealogical ancestry of a sample of genes from the present day.
- The Kolmogorov forward equation is a PDE describing the dynamics of the WF model in a large but finite population.
-

Lecture 5.8 Birth-Death Models

- Results from the appendix of Raup 1985

The Yule Model

The Yule process is a stochastic process proposed by George Yule in 1924 to describe the distribution of species among different taxonomic groups, known as clades.

In the context of ecology or biology, the Yule process is a branching process that models the evolution of species in a phylogenetic tree. The process assumes that new species can arise by speciation events, where an existing species splits into two new ones.

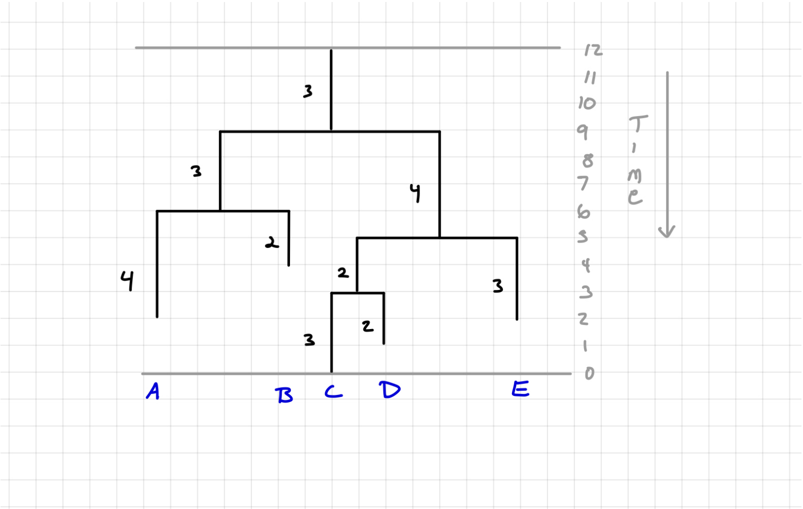
Let $N(t)$ be the number of species present in a clade at time t. Speciation occurs (species are 'born') at a constant per-capita rate λ. Then $N(t)$ is a continuous-time stochastic process known as the Yule process or "birth process".

Simulating a Phylogenetic Tree under the Yule Model

The key to the simulation here is the storage of data. We can represent a tree as a **cophenetic** matrix where element $C_{i,j}$ give the total amount of "shared ancestry between the lineages".

Example: 5.XX cophenetic matricies

What is the cophenetic matrix for the tree shown below?



To simulate a tree under the Yule process we simply need to simulate the corresponding cophenetic matrix.

Initialization: Initially the cophenetic matrix is:

Initially the time is $t = 0$

$$\mathbf{C}(0) = [0]$$

There is a single lineage/point that shares no ancestry with itself.

Iteration For ($t < t_{max}$):

- Calculate the total rate of speciation $\Lambda = \lambda * n$ where n is the number of lineages in $\mathbf{C}$.
- Draw the time until the next event $\Delta t \sim Exp(\Lambda)$. Update $t+ = \Delta t$
- Update the matrix by adding Δt to each diagonal element. This extends each existing branch by the amount Δt.
- Add a branching event
    - Choose a parent to branch at random
    - Copy and paste the parent's row and column and add t to the current diagonal. This adds the offspring which by definition shares t ancestry with itself.

Output: Return the cophenetic matrix

Example: 5.18 Yule Simulation

Python Lec_5.8

1. Simulate the Yule process with $\lambda = 1$ for $T = 1.5$ units of time. Sketch the corresponding tree.

2. Draw the "Lineage Through Time" plot for your simulation

The Yule simulations above have a behaviour similar to exponential population growth. How big should the tree be after T units of time and how should the number of lineages change through time?

Clade size in the Yule model

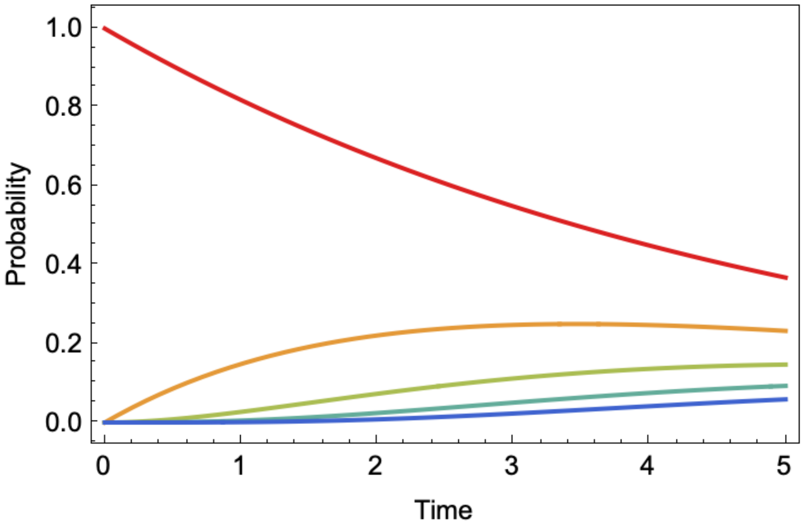
Specifically, suppose we start with 1 lineage at time $t = 0$ (the origin of the tree), how big is the tree at time $t = T$ later? To calculate this we consider the master equations for the probability that there are n lineages.

$$\frac{dP_n(t)}{dt} = -\lambda n P_n + \lambda(n-1)P_{n-1} \quad P_n(0) = \begin{cases} 0 & n \neq 1 \\ 1 & n = 1 \end{cases}$$

The solution to these ODEs are:

$$P(n,t) = e^{-n\lambda t}\left(e^{\lambda t} - 1\right)^{(n-1)}$$

The solutions for $n = 1$(red) to $n = 5$(blue) are shown below:



The Constant-Rate Birth-Death Model

The Yule model has one key assumption that deviates from our knowledge of the natural world, the fact that species can both speciate (be born) and go extinct (die). The simplest model that captures this process is the constant rate birth-death model in which birth events occur at a constant per-capita rate λ and die at a per-capita rate μ.

Simulating the Birth-Death Model

The simulation here is similar to that above in that we are creating a cophenetic matrix, except now we must consider the possibility of extinctions as well.

Iteration For $t < t_{max}$ or $n == 0$:

- Calculate the total rate of events $\Lambda = \lambda * n + \mu * n$ where n is the number of lineages in $\mathbf{C}$.
- Draw the time until the next event $\Delta t \sim Exp(\Lambda)$. Update $t+ = \Delta t$
- Update the matrix by adding Δt to each diagonal element. This extends each existing branch by the amount Δt.
- Choose the type if event: Speciation with probability $\frac{\lambda}{\lambda+\mu}$ and Extinction with probability $\frac{\mu}{\lambda+\mu}$
- If branching event
    - Choose a parent to branch at random
    - Copy and paste the parent's row and column and add t to the current diagonal. This adds the offspring which by definition shares t ancestry with itself.
- Else (extinction):
    - Choose a lineage to go extinct
    - Remove row and column corresponding to that lineage.

Output: Return the cophenetic matrix

Example: 5.19 Birth-Death Simulation

Python Lec_5.8

1. Simulate the Birth Death process with $\lambda = 1.5$ $\mu = 0.5$ for $T = 1.5$ units of time. Sketch the corresponding tree.

Probability of extinction in the BD model

To model the probability of extinction in the birth-death model we can use the theory of branching processes. Recall that we begin by specifying a offspring distribution, the probability of having k offspring. A lineage gives rise to two lineages if it speciates before going extinct $\frac{\lambda}{\lambda+\mu}$ and 0 offspring if it goes extinct before speciating.

$$\Pr(k) = \begin{cases} \frac{\lambda}{\lambda+\mu} & k = 2 \\ \frac{\mu}{\lambda+\mu} & k = 0 \end{cases}$$

Recall that we can then calculate the probability of extinction g as:

$$g = 1 * \frac{\mu}{\lambda + \mu} + g^2 * \frac{\lambda}{\lambda + \mu}$$

Solving for the roots of this equation we have:

$$\hat{g} = 1 \quad \hat{g} = \frac{\mu}{\lambda}$$

The first gives the probability of extinction in infinite time. The second gives the probability of extinction in finite time.

Discussion: does the probability of extinction in finite time make sense? How could you double-check it?