Markov Chains in Biology: Moran Model

Mathematical Population Genetics:
quantitatively understand - genetic diversity changes

Basic terminology:

- Gene: basic of unit of heredity. Refers to particular chromosomal locus
- Alleles: different versions of information encoded at the genetic locus.

Example: Pea seeds-wrinkled or round.
shape: 2 alleles

Outline of the Talk:

1) Setup - Moran Model - Results
2) Proof of the Results
3) Some Generalizations: of the Moran Model. K-allele model, selection, mutation
4) Wright-Fisher: setup, results, relation to the Moran Model

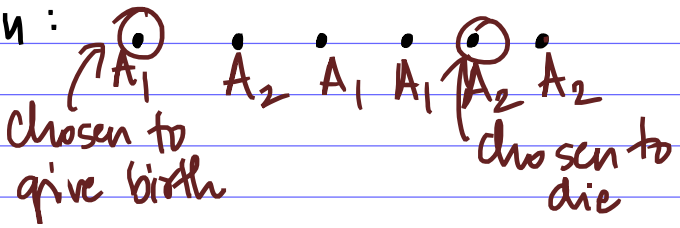
Section 1 : Setup and Results:

Finite population of size m
consisting of individuals of 2 types - A_1 and A_2

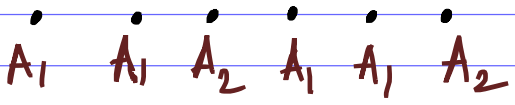
At time n:

- 1 individual is chosen randomly to give birth to a new individual of the same type
- 1 individual chosen randomly to die.

$m=6$: At time n :



New population:



X_n - number of individuals of type A_1 at time n

$\{X_n\}_{n\geq 0}$ is a Markov chain, $S=\{0,1,\dots,m\}$

Transition matrix P: $p_{00} = p_{mm} = 1$

When state $\{0\}$ is reached, all individuals are of the type A_2.
composition of the population will not change

When state $\{m\}$ is reached all individuals are of type A_1,

composition - unchanged.

Suppose $i \in \{1, 2, \dots, m-1\}$.

$P_{i,i-1} = \frac{i(m-i)}{m^2}$

Individual of type A_2 is chosen to give birth - w.p. $(m-i)/m$

Individual of type A_1 is chosen to die w.p. i/m.

$P_{i,i+1} = \frac{i(m-i)}{m^2}$

Individual of type A_1 chosen to give birth w.p. i/m

Individual of type A_2 die - w.p. $\frac{m-i}{m}$

$P_{ii} = \frac{i^2 + (m-i)^2}{m^2}$

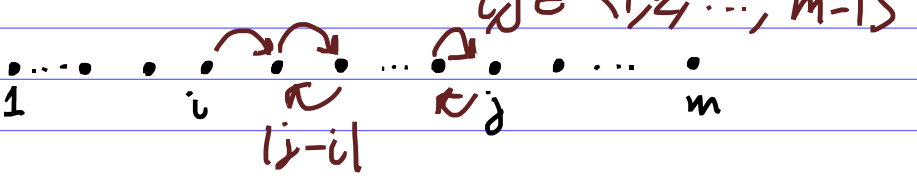
Either both individuals chosen are of type A_1 w.p. i^2/m^2

or both individuals chosen are of type A_2 w.p. $(m-i)^2/m^2$

Joint distribution:

$$P(X_0 = i_0, X_1 = i_1, \dots, X_n = i_n) = a_{i_0} P_{i_0 i_1} P_{i_1 i_2} \cdots P_{i_{n-1} i_n}$$

- Communicating classes: $\{0\}$, $\{m\}$, $\{1, 2, \dots, m-1\}$



- Recurrent states: $\{0\}$, $\{m\}$ $\quad P_{00} = P_{mm} = 1$

Transient states: $\{1, 2, \dots, m-1\}$

Genetic Interpretation:

Individuals of type A_1 and A_2 - interpreted as alleles of a particular gene.

This population has the following characteristics:

- haploid : each cell has 1 set of chromosomes A_1 or A_2
- monoecious : offsprings produced without mating

Hitting probability: of the state $\{0\}$
fixation probability of the allele A_2.

Hitting probability: of state $\{m\}$
fixation probability of the allele A_1.

Lemma 1: Fixation Probability:

In the Moran model, fixation probability of the allele A_1, when initially i copies are present, is given by:

$$P_i(X_n = m \text{ for some } n \geq 0) = \frac{i}{m}$$

$X_0 = i$ w.p. 1.

- How long does it take for the genetic diversity to disappear especially for a large populations.

Theorem 1: Mean Time to Fixation for a Large Population

$$T := \min\{n \geq 0, X_n \in \{0, m\}\}$$

Mean time to fixation, when there were i copies initially, is given by $E_i(T)$

$$\lim_{m \to \infty} \frac{E_i(T)}{\left\{\left(1-\frac{i}{m}\right)\log\left(1-\frac{i}{m}\right) + \frac{i}{m}\log\frac{i}{m}\right\} - m^2} = 1$$

Put $p = i/m$, $p \in (0,1)$.

$$E_{pm}(T) \approx -m^2\left((1-p)\log(1-p) + p\log p\right)$$

Term: $p\log p + (1-p)\log(1-p)$ entropy of Bernoulli(p) r.v.

For a discrete r.v. X, the entropy of X, denoted by $H(X)$, is defined as

$$H(X) = -\sum_{x \in \text{Range}(X)} P(x)\log(P(x))$$

$P(X)$: pmf of X

Entropy measures the randomness in a system

Section 2 : Proof of Results

Proof of Lemma 1 :

$$\tau_0 := \min\{n \geq 0 : X_n = 0\}$$
$$\tau_m := \min\{n \geq 0 : X_n = m\}$$
} stopping times

$$T = \min\{n \geq 0 : X_n \in \{0, m\}\}$$
$$T = \tau_0 \wedge \tau_m$$

Fact: $\{X_n\}_{n \geq 0}$ is a Martingale

Apply Optional Stopping Theorem

- $\mathbb{P}(T < \infty) = 1$ - $\{0\}, \{m\}$ only recurrent states
- $\{X_n\}_{n \geq 0}$ is a bounded Martingale.

$\Rightarrow \mathbb{E}(|X_T|) < \infty$ and $\mathbb{P}(X_n \mid T > n)\,\mathbb{P}(T > n) \to 0$ as $n \to \infty$

(see : Lecture on 23 March)

$$\mathbb{E}(X_T) = \mathbb{E}(X_0)$$

$$\mathbb{E}_i(X_T) = \mathbb{E}_i(X_0) = i$$

X_T can have 2 values : $\{0\}, \{m\}$

$$m\,\mathbb{P}_i(X_T = m) = i$$
$$\mathbb{P}_i(\tau_m < \tau_0) = i/m$$

Also, $\mathbb{P}_i(\tau_0 < \tau_m) + \mathbb{P}_i(\tau_m < \tau_0) = 1$]

$$\mathbb{P}_i(\tau_0 < \tau_m) = \frac{m - i}{m}$$

∎

Proof of Theorem 1:

Want to find $\mathbb{E}_i(T) =: k_i$

Fix $i \in \{1, 2, \cdots, m-1\}$

Suppose $j \in \{1, 2, \cdots, m-1\}$
mean time spent in the state j, starting at i, before absorption: k_i^j

$$S_j := \sum_{n=0}^{\infty} \mathbb{1}(X_n = j)$$

counts the amount of time spent j

$$k_i^j = \mathbb{E}_i(S_j)$$

Claim 1: $k_i^j = \delta_{ij} + p_{ii-1} k_{i-1}^j + p_{ii} k_i^j + p_{ii+1} k_{i+1}^j$

δ_{ij} - Kronecker Delta function

$$k_0^j = k_m^j = 0$$

This gives a recurrence relation in k_i^j.

Claim 2: Solution of the above recurrence relation is:

$$k_i^j = \begin{cases} i/j \; m & \text{for } i \le j \\ \frac{m-i}{m-j} \; m & \text{for } i > j \end{cases}$$

$$k_i = \sum_{j=1}^{m-1} k_i^j = m \left\{ \sum_{j=1}^{i} \frac{m-i}{m-j} + \sum_{j=i+1}^{m-1} \frac{i}{j} \right\}$$

Look at the case when m is large.
$p := i/m$. Then $p \in (0,1)$

$$k_{pm} = m\left\{\sum_{j=1}^{pm}\frac{m-pm}{m-j} + \sum_{j=pm+1}^{m-1}\frac{pm}{j}\right\}$$

$$= m^2\left\{(1-p)\sum_{j=1}^{pm}\frac{1}{m-j} + p\sum_{j=pm+1}^{m-1}\frac{1}{j}\right\}$$

Lemma 0.2:

$$\sum_{j=1}^{pm}\frac{1}{m-j} \approx -\log(1-p)$$

$$\sum_{j=pm+1}^{m-1}\frac{1}{j} \approx -\log p$$

$$k_{pm} = \mathbb{E}_{pm}(T) \approx -m^2\left((1-p)\log(1-p) + p\log p\right)$$

Proof of Claim 1:

$$S_j = \sum_{n=0}^{\infty}\mathbb{1}(X_n = j)$$

$$k_i^j = \mathbb{E}_i(S_j) = \mathbb{E}(S_j \mid X_0 = i)$$

$$= \sum_{n=0}^{\infty}\mathbb{E}(\mathbb{1}(X_n = j) \mid X_0 = i)$$

Indicator functions are non-negative r.v.s

$$= \delta_{ij} + \sum_{n=1}^{\infty}\mathbb{E}(\mathbb{1}(X_n = j) \mid X_0 = i)$$

$$= \delta_{ij} + \sum_{n=1}^{\infty}\sum_{l=0}^{m}\mathbb{E}(\mathbb{1}(X_n = j), X_1 = l \mid X_0 = i)$$

$$= \delta_{ij} + \sum_{n=1}^{\infty}\sum_{l=0}^{m}\mathbb{E}(\mathbb{1}(X_n = j) \mid X_1 = l, X_0 = i)\,\underline{P(X_1 = l \mid X_0 = i)}$$

$$= \delta_{ij} + \sum_{n=1}^{\infty}\sum_{l=0}^{m} p_{il}\,\mathbb{E}(\mathbb{1}(X_n = j) \mid X_1 = l, X_0 = i)$$

$$= \delta_{ij} + \sum_{l=0}^{m} p_{il} \sum_{n=1}^{\infty} \mathbb{E}(\mathbb{1}(X_n = j) \mid X_1 = l, X_0 = l)$$

$$= \delta_{ij} + \sum_{l=0}^{m} p_{il} \left(\delta_{jl} + \sum_{n=2}^{\infty} \mathbb{E}(\mathbb{1}(X_n = j) \mid X_1 = l) \right)$$

$$= \delta_{ij} + \sum_{l=0}^{m} p_{il} \, k_l^j$$

$i, j \in \{1, 2, \cdots, m-1\}$

$$k_i^j = \delta_{ij} + p_{i,i-1} k_{i-1}^j + p_{ii} k_i^j + p_{i,i+1} k_{i+1}^j$$

Section 3 : Some Generalizations :

K-allele Moran Model:

many genes have more than two alleles.

Example : ABO blood group system

Behaviour of population is studied by observing a set of Markov chains :

$$\{X_n^{(1)}, X_n^{(2)}, \cdots, X_n^{(k-1)}\}_{n \geq 0}$$

$X_n^{(i)}$: number of alleles of type i at the time n.

Selection: alleles - selective advantage - chosen for reproduction or to die.

Mutation: The alleles A_1 and A_2 can mutate to each other with positive probability.

The structure of the Markov chain changes.

$\{0\}, \{m\}$ - are no longer closed states.

irreducible Markov chain - recurrent.

Section 4: Wright-Fisher Model

Take a population of $m/2$ individuals where m is an even number.

Suppose the population is diploid, ie, each individual possesses two copies of each chromosomes - one from each parent.

Let the two alleles: A_1 and A_2.

Each gene looks like: A_1A_1, A_1A_2, A_2A_2.

- At time epoch n, next generation:

choosing randomly (with replacement) from the previous generation.

X_n: number of alleles of type A_1 at time n.

$\{X_n\}_{n\geq 0}$: Markov chain, $S=\{0,1,\dots,m\}$

Transition Matrix: $p_{00} = p_{mm} = 1$

Once $\{0\}$ is reached, A_2 - composition can no longer be changed.

$i \in \{1,2,\dots,m-1\}$

Given: $X_n = i$

$$X_{n+1} \mid X_n = i \sim Bin\left(m, \frac{i}{m}\right)$$

$$p_{ij} = \binom{m}{j}\left(\frac{i}{m}\right)^j\left(\frac{m-i}{m}\right)^{m-j} \quad \text{for } (i,j) \neq (0,0) \text{ and } (m,m)$$

Results:

- Communicating classes : $\{0\}$, $\{m\}$, $\{1, 2, \dots, m-1\}$

- Recurrent states: $\{0\}$, $\{m\}$ $p_{00} = p_{mm} = 1$
  Transient states: $\{1, 2, \dots, m-1\}$

- Hitting probability of the state $\{m\}$, when there are initially i copies of the allele, is given by:

$$h_i = P_i(X_n = m \text{ for some } n \geq 0) = \frac{i}{m}$$

- $\{X_n\}_{n \geq 0}$ is a Martingale.

- Theorem : The mean time for fixation of the alleles is given by $E_i(T)$

$$\lim_{m \to \infty} \frac{E_i(T)}{-2m\left(\left(1 - \frac{i}{m}\right)\log\left(1 - \frac{i}{m}\right) + \frac{i}{m}\log\frac{i}{m}\right)} = 1$$

$p := i/m$. $p \in (0, 1)$.

$$E_{pm}(T) \approx -2m\underbrace{\left((1-p)\log(1-p) + p(\log p)\right)}$$

differs only by a factor of $m/2$ as compared to the Moran model.

Both the Wright-Fisher model and the Moran model have the same limiting distribution, Kingman's coalescent, but each model follows a different.

(see Theorem 1.30 and Theorem 4.1 in the book Probability Models in DNA Sequence Evolution by Richard Durrett).