Computational Systems Biology: Biology X

Bud Mishra

Room 1002, 715 Broadway, Courant Institute, NYU, New York, USA

Human Population Genomics

Outline

"Damn the Human Genomes. Small initial populations; genes too distant; pestered with transposons; feeble contrivance; could make a better one myself."

–Lord Jefferey (badly paraphrased)

Models in Population Genetics

Population Models

These are models to describe the evolution of allele frequencies...

There are two classical models

Wright-Fisher model

Moran model

Outline

Wright-Fisher model

- Assume a simple haploid model; it consists of a population of $2N$ genes (or alternatively – N diploid organisms) of random reproduction, with each haploid possessing either allele A_1 or allele A_2.
- Initially, we may disregard mutation as well as selective forces.
- At each time-step, a gene (allele) reproduces some number of offspring (which are the exact copies of itself) and dies immediately after that; thus has a life-span of only one generation.
- The process, modeled thus, describes how the genes get transmitted from one generation to the next.

Markov Model

- Process of birth and death in the population remains hidden. The only observable is the frequency of alleles changing from generation to generation. The allele frequency of the next generation is governed only by a genetic drift.

Definition (Genetic drift)

It is defined as a force that reduces heterozygosity by the random loss of alleles.

- Focus on the frequency of allele A_1 in the population of $2N$ haploids. Think of this process as changing from one generation to the next in terms of a Markov Chain, where the state X of the chain corresponds to the number of haploids (genes) of type A_1.

- In any generation X takes one of the values 0, 1, . . . , $2N$, which constitutes a *state space*. Denote the value taken by X in generation t as X_t.
- The model assumes that genes for the generation $t + 1$ are derived by sampling with replacement from the genes of generation t. Thus, the make up of the next generation is determined by $2N$ independent Bernoulli trials so that X_t is a binomial random variable.

- Let the initial generation consist of i genes of type A_1 and $2N - i$ genes of type A_2. Then we define a probability of success (resulting in allele A_1) p_i and a probability of failure q_i (resulting in allele A_2) for each Bernoulli trial as

$$p_i = \frac{i}{2N} \qquad q_i = 1 - \frac{i}{2N}$$

- The process generates a Markov Chain $\{X_n\}$, where X_n is the number of A_1 genes in the n^{th} generation, among a constant population size of $2N$ individuals. Basically, X_{t+1} is a binomial random variable with index $2N$ and parameter (probability of success) $X_t/2N$.
- Observe that the transition probabilities from $X_t = i$ to $X_{t+1} = j$ for this Markov Chain are computed according to the binomial distribution as

$$\begin{aligned} P(X_{t+1} = j | X_t = i) &= p_{ij} = \binom{2N}{j} p_i^j q_i^{2N-j} \\ &= \binom{2N}{j} (i/2N)^j \{1 - (i/2N)\}^{2N-j} \end{aligned}$$

- *The states* 0 *and* $2N$ *are completely absorbing.*
- In other words, no matter what the value of X_0 is, eventually X_t will take the value 0 or $2N$.Once this happens, X will stay in that state forever. In the case of $X_t = 0$, the population will consist only of A_2 genes, while in the case of $X_t = 2N$ the population will be purely A_1-gene population.

Absorption probability in Wright-Fisher model

- In this model, eventually, the population attains fixation; at that point, it is composed of only A_1-genes ($X_t = 2N$) or A_2-genes ($X_t = 0$). That is, with probability one, either of the absorbing states (either 0 or $2N$) is eventually entered (and this is true for both Wright-Fisher and Moran models).
- Thus, for $0 < j < 2N$,

$$\lim_{t \to \infty} P(X_t = j) = 0.$$

Absorption at Zero

- Probability of extinction (absorption at 0) of a gene, given that it started with i copies

$$lim_{n\to\infty} P(X_n = 0 | X_0 = i).$$

- Note that

$$\begin{aligned} E(X_n) &= E[E(X_n|X_{n-1})] = E(X_{n-1}) = E(X_{n-2}) \\ &= \cdots = E(X_0) = i. \end{aligned}$$

This property is called the *constancy of expectation*. It is also true for Moran model.

- Note further that

$$E(X_n) = 0 \cdot u_{i,0} + 2N \cdot (1 - u_{i,0}).$$

- Since $lim_{n \to \infty} E(X_n) = i$, we have

$$i = 0 \cdot u_{i,0} + 2N \cdot (1 - u_{i,0}),$$

and therefore

$$u_{i,0} = \frac{2N - i}{2N}.$$

Absorption at $2N$

- In an identical manner, we may calculate the probability that A_1 eventually becomes fixed in the population (absorption at $2N$)...

$$i = 0 \times (1 - u_{i,2N}) + 2N \times u_{i,2N}$$

Thus

$$u_{i,2N} = \frac{i}{2N}$$

- Eventually every gene in the population is descended from one unique gene which appeared in generation zero. The probability that such a gene (allele) is A_1 is simply the initial fraction of A_1 alleles, namely $i/2N$, and this also must be a fixation probability of allele A_1.

Absorption starting from a Single A_1 Allele

- In a population of pure A_2 alleles a single new mutant A_1 allele (gene) arises.
- Since it is assumed that there are no more new mutations, we are starting with a population with one A_1 allele and $2N-1$ A_2 alleles.
- Thus, the probability of fixation for this allele is

$$u_{1,2N} = \frac{i}{2N} = \frac{1}{2N}.$$

- Symmetrically, the probability that the allele is lost is $1 - 1/2N$.

Mean Time till Absorption in Wright-Fisher Model

- In general, computationally expensive; but can be approximated.
- Easy case: Mean time until absorption starting with one allele of type A_1
  ...(before the mutant is lost or before the mutant is fixed).
- A cute trick... Just add up the expected number of visits to a state j along the path to absorption, starting from state $X_0 = 1$. ... Denote the mean number of generations to absorption in 0 or $2N$, given that the population started with just one allele A_1, as $\bar{t}_1$.

- Summing up the expected number of such visits for all j, avoiding the absorbing states: 0 and $2N$:

$$\bar{t}_1 = \sum_{j=1}^{2N-1} \bar{t}_{1,j},$$

where $\bar{t}_{1,j}$ is the mean number of times when the number of A_1 alleles assumes the value of j (i.e., the system is in state j) before reaching either 0 or $2N$. Thus

$$\bar{t}_{1,j} \approx \frac{2}{j},$$

starting at $i = 1$.

- Since $\sum_{i=1}^{N} \frac{1}{i} = \ln(N) + \gamma$ where γ is the Euler's constant (0.5772...), we have

$$\bar{t}_1 = \sum_{j=1}^{2N-1} \bar{t}_{1,j} = \sum_{j=1}^{2N-1} \frac{2}{j} = 2 \sum_{j=1}^{2N-1} \frac{1}{j}$$

$$= 2(\ln(2N-1) + \gamma)$$

Approximating Mean Time until Absorption

- A simple approximation for $\bar{t}_i$. Start from state i and in the first step reach some intermediate step k.
- Define $M = 2N$, $i/M = x$, $k/M = x + \delta x$, and $\bar{t}_i = \bar{t}(x)$.

$$\bar{t}_i = \sum_{k=0}^{M} p_{ik} \bar{t}_k + 1$$

as

$$\begin{aligned} \bar{t}(x) &= \sum P\{x \to x + \delta x\} \bar{t}(x + \delta x) + 1 \\ &= E\{\bar{t}(x + \delta x)\} + 1 \end{aligned} \tag{1}$$

- Assuming that $\bar{t}(x)$ is a twice differentiable function of a continuous variable x, we can use Taylor series to approximate the above quantity.

- The Taylor series states that

$$f(y) = \sum_{n=0}^{\infty} \frac{f^{(n)}(a)}{n!}(y-a)^n \tag{2}$$
$$= \frac{f(a)}{0!}(y-a)^0 + \frac{\{f(a)\}'}{1!}(y-a)^1 + \frac{\{f(a)\}''}{2!}(y-a)^2 + \ldots$$
$$= f(a) + \{f(a)\}'(y-a) + \frac{1}{2}\{f(a)\}''(y-a)^2 + \ldots,$$

- We re-write $\bar{t}(x)$ by applying Taylor's series

$$\bar{t}(x) = E\{\bar{t}(x+\delta x)\} + 1 \tag{3}$$
$$\approx \bar{t}(x) + E(\delta x)\{\bar{t}(x)\}' + \frac{1}{2}E(\delta x)^2\{\bar{t}(x)\}'' + 1,$$

- All expectations are conditional on x. Since the expectation of a binomial random variable $Y \sim B(n, p)$ is $E(Y) = np$,

$$E(x + \delta x) = E(j/M) = \frac{E(j)}{M} = \frac{M \cdot \frac{i}{M}}{M} = \frac{i}{M},$$

  where $p = \frac{i}{M}$.
- Since we know that $E(X_n | X_{n-1} = i) = i$, and since $x = \frac{i}{M}$ and $E(x) = x$, $E(\delta x) = 0$. As a result the term $E(\delta x)\{\bar{t}(x)\}' = 0$.

- Calculating $E(\delta x)^2$: In our case, the $E(\delta x)^2 = Var(\delta x)$ since $[E(\delta x)]^2 = 0$. The variance of the binomial r.v.

$$\begin{aligned} Var(x + \delta x) &= Var(j/2N) = \frac{Var(j)}{4N^2} \\ &= \frac{2N\frac{i}{2N}(1 - \frac{i}{2N})}{4N^2} = \frac{x(1-x)}{2N} \end{aligned}$$

.

- Hence,

$$\bar{t}(x) \approx \bar{t}(x) + \frac{1}{2}\frac{x(1-x)}{2N}\{\bar{t}(x)\}'' + 1$$

$$-1 \approx \frac{1}{2}\frac{x(1-x)}{2N}\{\bar{t}(x)\}''$$

$$-4N \approx x(1-x)\{\bar{t}(x)\}''$$

- The solution to this equation, subject to the boundary conditions $\bar{t}(0) = \bar{t}(1) = 0$ is

$$\begin{aligned} \bar{t}(x) &= \int\int -4N\frac{1}{x(1-x)} \qquad (4) \\ &= -4N\int\int\left(\frac{1}{x}+\frac{1}{1-x}\right) \\ &= -4N\int(\ln(x)+\ln(1-x)) \\ &\approx -4N\{x\ln x+(1-x)\ln(1-x)\} \end{aligned}$$

where $x = \frac{i}{2N}$, the initial frequency of allele A_1.

Diffusion Approximation to the Mean Absorption Time

- Starting with just one A_1 allele $x = \frac{1}{2N}$, the mean time to absorption is

$$\bar{t}\left(\frac{1}{2N}\right) = -4N\left\{\frac{1}{2N}\ln\left(\frac{1}{2N}\right) + \left(1-\frac{1}{2N}\right)\ln\left(1-\frac{1}{2N}\right)\right\}$$
$$\approx 2 + 2\ln 2N. \quad (5)$$

- When $x = \frac{1}{2}$,

$$\bar{t}\left(\frac{1}{2}\right) \approx 2.8N$$

- For equal initial frequencies ($x = \frac{1}{2}$), the mean time is relatively long.

Outline

Moran Model

- In each generation of the Moran model, one gene is chosen at random to give 2 offspring and one gene is chosen to die (all other genes survive to the next generation).
- In contrast to Wright-Fisher model, Moran model has overlapping generations. This model is also known as a birth-and-death model.
- We still consider a constant population size of $2N$ haploids, each of which has either allele A_1 or allele A_2. Let us (for now) ignore mutation or selection pressures.

- Define X to be a random variable, representing the number of alleles of type A_1 in the population. Question: What are the transition probabilities for the implied Markov chain?
- Suppose that in population at t (which corresponds to state X_t in underlying Markov chain), the number of alleles A_1 is i. Then in population $t+1$, the number of alleles A_1 can be either $(j = i-1)$, $(j = i+1)$, or $j = i$.

- The system can go from i to $i-1$ if A_2 is chosen to reproduce 2 offspring and A_1 is chosen to die; similarly, it can go from i to $i+1$, if A_2 is chosen to die and A_1 is chosen to reproduce:

$$p_{i,i-1} = \left(\frac{2N-i}{2N}\right)\left(\frac{i}{2N}\right) \quad \& \quad p_{i,i+1} = \left(\frac{2N-i}{2N}\right)\left(\frac{i}{2N}\right)$$

- And for going from i to i, it takes either A_1 to reproduce and die or A_2 to reproduce and die:

$$p_{i,i} = \left(\frac{2N-i}{2N}\cdot\frac{2N-i}{2N}\right) + \left(\frac{i}{2N}\cdot\frac{i}{2N}\right) = \frac{i^2+(2N-i)^2}{(2N)^2}$$

- Note that $p_{ij} = 0$ for all other values of j.

Properties of a Continuant matrix in Moran model.

- In Moran model, the transition probability matrix is a Continuant... $p_{ij} = 0$ iff $|i - j| > 1$.
- Using standard Continuant matrix theory, calculate explicitly the probability of fixation and mean time to absorption ...
- A "birth-and-death" process model (a special case of Continuous-time Markov process) ...When a birth occurs, the state i goes to state $i + 1$, defined by the birth rate $p_{i,i+1} = \lambda_i$. ...When a death occurs, the process goes from state i to state $i - 1$, defined by the death rate $p_{i,i-1} = \mu_i$.

- Define $\rho_0 = 1$ &

$$\rho_i = \frac{\mu_1 \mu_2 \dots \mu_i}{\lambda_1 \lambda_2 \dots \lambda_i},$$

- $M = 0$ and $M = 2N$ are both absorbing states; the probability of absorption in either of them becomes

$$u_i = \sum_{k=0}^{i-1} \rho_k \Big/ \sum_{k=0}^{M-1} \rho_k$$

- Thus the mean number of times the system is in state j given that it started in state i is

$$\bar{t}_{ij} = \frac{(1-u_i)\sum_{k=0}^{j-1}\rho_k}{\rho_{j-1}\mu_j}, (j = 1, \ldots, i) \quad \&$$

$$\bar{t}_{ij} = \frac{u_i \sum_{k=j}^{M-1}\rho_k}{\rho_j \lambda_j}, (j = i+1, \ldots, M-1)$$

-

$$\bar{t}_i = \sum_{j=1}^{M-1} \bar{t}_{ij} = \sum_{j=1}^{i} \frac{(1-u_i)\sum_{k=0}^{j-1}\rho_k}{\rho_{j-1}\mu_j} + \sum_{j=i+1}^{M-1} \frac{u_i \sum_{k=j}^{M-1}\rho_k}{\rho_j \lambda_j}$$

Probability of fixation in Moran model

- Thus, in the Moran model

$$\lambda_i = \mu_i = i(2N - i)/(2N)^2$$

so that

$$\rho_i = \frac{\mu_1}{\lambda_1}\frac{\mu_2}{\lambda_2}\cdots\frac{\mu_i}{\lambda_i} = 1$$

for $i = 0, 1, \ldots, 2N$.

- It can be shown that, similarly to Wright-Fisher model, $E(X_t) = i$.
- Putting all together... the probability of fixation (given that we started with i copies of A_1) is

$$u_i = \frac{i}{2N}$$

given that $M = 2N$.

Expected Absorption Time: $j \leq i$

- Note that $\rho_i = 1$; we can derive the following:
- First, for $j = 1, \ldots, i$

$$\begin{aligned}
\bar{t}_{ij} &= \frac{(1-u_i)\sum_{k=0}^{j-1}\rho_k}{\rho_{j-1}\mu_j} \\
&= \frac{(1-\frac{i}{2N})\sum_{k=0}^{j-1}1}{1\frac{2N-j}{2N}\frac{j}{2N}} \\
&= \frac{(\frac{2N-i}{2N})j\cdot 2N\cdot 2N}{(2N-j)j} \\
&= \frac{(2N-i)\cdot 2N}{2N-j},
\end{aligned} \tag{6}$$

Expected Absorption Time: $j > i$

- Next, similarly, for $j = i+1, \ldots, M-1$

$$\begin{aligned}
\bar{t}_{ij} &= \frac{u_i \sum_{k=j}^{M-1} \rho_k}{\rho_j \lambda_j} \\
&= \frac{\frac{i}{2N} \sum_{k=j}^{M-1} 1}{1 \frac{2N-j}{2N} \frac{j}{2N}} \\
&= \frac{\frac{i}{2N}(2N-j) \cdot 2N \cdot 2N}{(2N-j)j} \\
&= \frac{i \cdot 2N}{j}.
\end{aligned} \tag{7}$$

- The expected time to absorption can be expressed as

$$\begin{aligned}
\bar{t}_i &= \sum_{j=1}^{M-1} \bar{t}_{ij} \\
&= \sum_{j=1}^{i} \frac{(1-u_i)\sum_{k=0}^{j-1}\rho_k}{\rho_{j-1}\mu_j} + \sum_{j=i+1}^{M-1} \frac{u_i \sum_{k=j}^{M-1}\rho_k}{\rho_j \lambda_j} \\
&= \sum_{j=1}^{i} \frac{(2N-i)\cdot 2N}{2N-j} + \sum_{j=i+1}^{M-1} \frac{i\cdot 2N}{j} \\
&= (2N-i)2N \sum_{j=1}^{i} \frac{1}{2N-j} + 2Ni \sum_{j=i+1}^{M-1} \frac{1}{j} \qquad (8)
\end{aligned}$$

Outline

Other Models

- Coalescent and Phylogenetic Trees
- Ancestral States and MRCA (Most Recent Common Ancestor)
- Population Changes in the Coalescent
- Population Bottlenecks
- General Models: Finite Island Models, Models of Division, Non-Equilibrium Models
- Coalescent with Selection
- Coalescent with Recombination
- Parameter Estimation from Data
- LD mapping and the Coalescent
- Human Evolution, Migration and Population Structure

Coalescence in the Island Model

- Model of Population subdivision and migration [Wright(1931)]
- Population is divided into *D demes* each of size *N* haploid individuals...
- Each island accepts a fraction *m* migrations every generation
- It does not take into account explicit geography... Nor the nature of migration (e.g., urbanization, forced evacuation, etc.)... “Scatter-Gather” nature of migration affecting structure population...

[End of Lecture #7]

THE END