Computational Systems Biology: Biology X

Bud Mishra

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

Human Population Genomics

Outline

1. Recapitulation: Wright-Fisher & Moran models

2. Coalescence

"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)

Outline

1. Recapitulation: Wright-Fisher & Moran models
2. Coalescence

Wright-Fisher model

- Model of population for genealogical relationship among genes — Wright (1931) and Fisher (1930).
- Idealized haploid model of reproduction: Model of transmission of genes from one generation to the next in a population of fixed size; population of $2N$ genes, corresponding to N diploid or $2N$ haploid individuals.
- Each of the genes of generation $t+1$ are obtained by copying the gene of a random individual from generation t; this process is repeated until $2N$ genes have been sampled to create the population at $t+1$.
- A gene in generation t might not have any descendant in generation $t+1$ and thus its lineage dies out.

Moran model

- An alternative model to Wright-Fisher — Moran (1958)
- Moran model allows overlapping generations
- The population has $2N$ haploid individuals or genes
- A new generation is created from the previous one by sampling randomly randomly to give birth to a new gene, and one gene to die: The gene that dies is distinct from the one that gives birth. Population size remains fixed.
- The Moran model rules out the possibility of multiple coalescent events in the same generation (i.e., no more than two genes share the same common ancestor in the previous generation).

- Thus, one out of $\binom{2N}{2}$ possible pairs has the desired coalescent property, Thus the natural time scale is in units of $N(2N-1)$ Moran-generations, rather than in units of $2N$ Wright-Fisher genrations.
- After adjusting for the differences in time scales, the two models have approximately equivalent coalescence and fixation properties.

Assumptions of the Wright-Fisher Model

- **Discrete and non-overlapping generations**: For humans, a generation (from conception to reproduction) is assumed to be about 25 years.
- **Haploid individuals vs. two subpopulation**: Note that in practice, generation time differs for males and females, e.g., 30 vs. 20 years. If the selection does not involve heterosis, the difference has little quantitative consequence.
- **The population size is constant**: Population bottleneck effects not accounted for.

Assumptions of the Wright-Fisher Model

- **All individuals are equally fit**: Presence and strength of natural selection is ignored.
- **The population has no geographical or social structure**: It is a hard assumption to relax; but very important in modeling mechanism of reproduction in a real population.
- **The genes do not recombine within the population**: Mitochondria and Y chromosomes are possible exceptions... Must be modeled by an ancestral recombination graph.

Number of Descendants

- Number of descendants of a particular gene, i, in generation t: A stochastic variable.
- Let v_i be the number of descendants of gene i in generation t... $1 \leq i \leq 2N$.

$$Pr(v_i = k) = \binom{2N}{k} \left(\frac{1}{2N}\right)^k \left(1 - \frac{1}{2N}\right)^{2N-k} \approx \frac{1}{k!}e^{-1}.$$

- This is a binomial distribution $\mathrm{Bin}(m, p)$ ($m = 2N$; $p = 1/2N$) with a Poisson approximation $\mathrm{Poisson}(1)$.

- The moment generating function is $\psi(t) = \left[1 + \frac{(e^t - 1)}{2N}\right]^{2N}$, and for v_i, its mean is 1 and variance is

$$2N \frac{1}{2N}\left(1 - \frac{1}{2N}\right) = 1 - \frac{1}{2N}.$$

- If mean number had deviated from one, the population would grow without bound, or shrink to extinction.
- The covariance of the off-spring number for two genes i and j is

$$\mathrm{Cov}(v_i, v_j) = E(v_i v_j) - E(v_i)E(v_j) = -\frac{1}{2N}.$$

- The correlation coefficient is

$$\mathrm{Corr}(v_i, v_j) = \frac{\mathrm{Cov}(v_i, v_j)}{\sqrt{\mathrm{Var}(v_i)\mathrm{Var}(v_j)}} = -\frac{1}{2N - 1}.$$

Covariance

- A negative covariance is expected because if gene i leaves many descendants in next generation, then gene i is more likely to leave few.
- However, v_i and v_j are almost independent of each other for large $2N$.
- Note that the probability that a gene has no immediate descendant is $Pr(v_i = 0) = e^{-1}$. Thus approximately 0.63 fraction of all genes have descendants.
- In a few generations (i.e., relative to $2N$) a randomly mating population descends from a small number of genes.

Descendants

- If d_j denotes the probability that a gene in generation j leaves no descendant in the present generation, then $d_1 = e^{-1} \approx 0.37$. Furthermore,

$$d_j = \sum_{k=0}^{\infty} \frac{1}{k!} e^{-1} (d_{j-1})^k = e^{d_{j-1}-1}, \qquad \text{for } j > 1.$$

- For example, $d_{10} = 0.85$ and $d_{50} = 0.96$.
- An entire population of size $2N = 10,000$ descends from approximately $2N(1 - d_{50}) = 400$ genes 50 generations ago.

Outline

1. Recapitulation: Wright-Fisher & Moran models
2. Coalescence

Coalescence of a Sample of Two Genes

- What is the distribution of the waiting time until the MRCA (Most Recent Common Ancestor) of two genes sampled in a model with $2N$ genes?
- (a) The probability p that these two genes find an ancestor in the first generation back in time is $p = \frac{1}{2N}$ the first gene chooses its parent freely, the second must choose the same parent out of $2N$ possibilities; (b) The probability q that the two genes have different ancestors is therefore $q = 1 - \frac{1}{2N}$.
- The probability that the two genes finds a common ancestor exactly j generations back is

$$Pr(T_2 = j) = q^{j-1}p = \left(1 - \frac{1}{2N}\right)^{j-1} \frac{1}{2N}.$$

- Note

$$\begin{aligned} Pr(T_2 \geq j) &= q^{j-1}p[1+q+q^2+\cdots] \\ &= q^{j-1} \approx 1 - e^{-(j-1)/2N}. \end{aligned}$$

- Thus

$$\begin{aligned} Pr(T_2 \leq j) &= p[1+q+q^2+\cdots q^{j-1}] \\ &= 1 - q^j \approx 1 - e^{-j/2N}. \end{aligned}$$

- Note that these models assume a *Markov Property*: That is the probability of an event (such as *coalescence*) depends on the present state of the population — **The process has no memory of events prior to the present.**
- It also implicitly assumes that **the number of offsprings is distributed as a Poisson process with parameter** 1. In reality the mean or variance of the number of offspring may deviate from the expected value 1 (with population bottlenecks, etc.) They result in significant deviation from the predicted model.

Statistics

- Thus $T_2 \sim \text{Geo}(1/2N)$ is geometrically distributed with parameter $p = \frac{1}{2N}$. Hence, it has mean and variance

$$\begin{aligned}\text{Mean} = E(T_2) &= \frac{1}{p} = 2N \\ \text{Variance} = Var(T_2) &= \frac{1-p}{p^2} = 2N(2N-1).\end{aligned}$$

- Thus the expected time until a MRCA is the same as the number of genes in the population.

Coalescence of a Sample of n Genes

- The waiting time for $k (\leq n)$ genes to have less than k ancestral lineages: The probability that k genes have exactly k different ancestors in the previous generation is

$$\begin{aligned}\frac{(2N-1)}{2N}\frac{(2N-2)}{2N} \quad \cdots \quad \frac{(2N-k+1)}{2N} &= \prod_{i=1}^{k-1}\left(1-\frac{i}{2N}\right) \\ &= 1-\binom{k}{2}\frac{1+o(1)}{2N}.\end{aligned}$$

- Thus, as before, we have

$$Pr(T_k = j) \approx \left\{1 - \binom{k}{2}\frac{1}{2N}\right\}^{j-1}\binom{k}{2}\frac{1}{2N}.$$

- Thus T_k has approximately a geometric distribution with parameter $\binom{k}{2}/(2N)$. Note that the times $T_2, \cdots, T_n$ are independent.

Properties of Geometric Distributions

- Assume that $t_2 > t_1$. Then

$$Pr(T > t_2 | T > t_1) = Pr(T > t_2 - t_1).$$

- Let S and T be two independent geometrically distributed random variables. $S \sim Geo(p)$ and $T \sim Geo(p')$, then

$$\min(S, T) \sim Geo(p + p' - pp').$$

Properties of Exponential Distributions

- Assume that $t_2 > t_1$, and $V \sim Exp(a)$ and $U \sim Exp(b)$ are two independent exponentially distributed random variables. Then

$$\begin{aligned}
Pr(U > t_2 | U > t_1) &= Pr(T > t_2 - t_1) \\
E(V) &= \frac{1}{a} \qquad Var(V) = \frac{1}{a^2} \\
E(U) &= \frac{1}{b} \qquad Var(U) = \frac{1}{b^2} \\
Pr(v < U) &= \frac{a}{a+b}, \text{ and} \\
\min(U, V) &\sim Exp(a+b)
\end{aligned}$$

Continuous Time Approximation

- One unit of time corresponds to the average time for two genes to find a common ancestor: $E(T_2) = 2N$ generations. Time is scaled by a factor of $2N$ (or N or in some cases, $4N$).
- Coalescent becomes independent of the population size. **The structure of the coalescent process is the same for any population as lon as the sample size is small relative to population size** $2N$.

$$n \ll 2N.$$

  Only the time scale differs between populations when $2N$ varies.

Rescaling Time

- Let $t = j/(2N)$, where j is time measured in generations. $j = 2Nt$. The waiting time, T_k^c, in the continuos representation (for k genes to have $k - 1$ ancestors) is exponentially distributed $T_k^c \sim Exp(\binom{k}{2})$.

$$Pr(T_k^c \leq t) = 1 - e^{-\binom{k}{2}t}.$$

Stochastic Algorithm to Sample Genealogies for n Genes

- **Algorithm**
  1. Start with $k = n$ genes. Repeat until $k = 1$:
     1. Simulate the waiting time T_k^c to the next event $T_k^c \sim Exp(\binom{k}{2})$.
     2. Choose a random pair (i, j) with $1 \leq i < j \leq k$ uniformly from the $\binom{k}{2}$ possible pairs.
     3. Merge i and j into one gene and decrease the sample size by one: $k \mapsto k - 1$.

Effective Population Size

- Most real populations show some form of reproductive structure: either due to geological proximity of individuals or due to social constraints. Also, the number of descendants of a gene in one generation does not follow the Poisson distribution with intensity one.
- For a real population, the population size of the haploid Wright-Fisher that “best approximates”the real population is called the effective population size N_e. One could choose one of the following two:

$$N_e^{(i)} = \frac{1}{2Pr(T_2 = 1)}, \quad \text{or} \quad N_e^{(t)} = \frac{E(T_2)}{2}.$$

- $N_e^{(i)}$ (*inbreeding effective population size*) relates to the immediate past, where as $N_e^{(t)}$ relates to the number of generations until an MRCA is found.
- For the haploid Wright-Fisher model, both definitions agree $N_e^{(i)} = N_e^{(t)} = N$, since

$$Pr(T_2 = 1) = \frac{1}{2N}, \quad \text{and} \quad E(T_2) = 2N.$$

Diploid Model

- In the diploid model with $N_f = cN$ females and $N_m = (1-c)N$ males:

$$Pr(T_2 = 1) = \left(1 - \frac{1}{2N}\right)\frac{N}{8N_f N_m}.$$

- Hence

$$N_e \approx 4c(1-c)N.$$

- There are other robust ways of defining effective population size: but the differences are minor.

Mutation

Three interesting models:

- The infinite alleles model — Kimura and Crow 1964
- The infinite sites model — Kimura 1969
- The finite sites model - Jukes and Cantor 1969

Mutations are assumed to be selectively neutral. Thus the mutation process can be separated from the genealogical process.

- In the absence of **selection**, the mutational process and the transmission of genes from one generation to the next are independent processes.
- Thus a sample configuration or *n* genes can be simulated using a two step procedure:
    1. Simulate the genealogy of *n* genes;
    2. Add mutations to the genealogy according to the **chosen** model.

The Wright-Fisher Model with Mutation

- Impose a process of mutation on top of the process of reproduction.
- Each gene chosen to be passed on is subject to a mutation with probability u. [[With probability $1-u$ the gene is copied without modification to the offspring, and with probability u it mutates.]]
- If we follow a lineage from the present time to the past, then with probability u the parental gene in generation t differs from the offspring gene at time $t+1$.
- The probability that a lineage experiences the first mutation j generations back is

$$Pr(T_M = j) = u(1-u)^{j-1} \approx \frac{u}{u-1} e^{-uj}.$$

Continuous Approximation

- If time is measured in units of $2N$ generations (like in coalescence) then

  $$Pr(T_M \leq j) = 1 - (1 - u)^j \approx 1 - e^{-\theta t/2} = Pr(T_M^c \leq t),$$

  where $t = j/(2N)$, $\theta = 4Nu$ and T_M^c is the time in $2N$ (assumed large) generations units.
- The parameter θ is called the *population mutation rate* or the *scaled mutation rate*. It also tells us about how fixation and mutations work against each other...

$n > 2$ Lineages

- Consider n disjoint lineages. The time until the first mutation event in any of the n lineages is exponentially distributed with parameter $n\theta/2$.
- If we wait for mutation events of coalescence events then the parameter of the exponentially distributed waiting time is the sum of the two parameters, which is

$$\binom{n}{2} + \frac{n\theta}{2} = \frac{n(n-1+\theta)}{2}.$$

- Whether the first event is a coalescence or a mutation is determined by a Bernoulli trial:
- With probability

$$\frac{\binom{n}{2}}{\binom{n}{2} + \frac{n\theta}{2}} = \frac{n-1}{n-1+\theta},$$

  the event is a coalescence; and
- With probability

$$\frac{\theta}{n-1+\theta},$$

  it is a mutation.

Stochastic Algorithm to Sample Genealogies with Mutations

- **Algorithm**
  1. Start with $k = n$ genes (sample size). Repeat until $k = 1$:
     1. Simulate the waiting time T_k^c to the next event $T_k^c \sim Exp(k(k-1+\theta)/2)$.
     2. With probability $(k-1)/(k-1+\theta)$ the event is coalescence, and with probability $\theta/(k-1+\theta)$ the event is mutation.
     3. **Case Coalescence**: Choose a random pair (i,j) with $1 \leq i < j \leq k$ uniformly from the $\binom{k}{2}$ possible pairs. Merge i and j into one gene and decrease the sample size by one: $k \mapsto k-1$.
     4. **Case Mutation**: Choose a lineage at random to leave. The sample size k remains unchanged.

[End of Lecture #10]

THE END