Evolutionary Dynamics Homework 5

Minghang Li

October 31, 2024 02:10 (America/New_York)

Problem 1: Pathways of carcinogenesis

> Consider three independent mutations $\{1, 2, 3\}$. Each mutation occurs after an exponentially distributed waiting time $T_i \sim \exp(\lambda i), i = 1, 2, 3$.

(a) What is the probability for the path $P = 3 \to 1 \to 2$?

Since the mutations are independent, let Exit_j denote the set of all possible mutations in step j, then we have:

- $\text{Exit}_1 = \{1, 2, 3\}$.
- $\text{Exit}_2 = \{1, 2\}$ (as 3 has already occurred).
- $\text{Exit}_3 = \{2\}$.

$$\begin{aligned}\text{Prob}(P) &= \prod_{j=1}^{k} \frac{\lambda_{i_j}}{\sum_{i \in \text{Exit}_i} \lambda_i} \\ &= \frac{\lambda_3}{\lambda_1 + \lambda_2 + \lambda_3} \cdot \frac{\lambda_1}{\lambda_1 + \lambda_2} \cdot \frac{\lambda_2}{\lambda_2} \\ &= \frac{\lambda_1 \cdot \lambda_3}{(\lambda_1 + \lambda_2)(\lambda_1 + \lambda_2 + \lambda_3)}\end{aligned}$$

(b) Assume cancer arises if any two of the three genes are mutated. How many possible genotypes are there? How many pathways? Compute the expected waiting time until any two out of three genes are mutated.

If cancer arises if any two of the three genes are mutated, then the possible genotypes are

$$\text{number of genotypes} = \binom{3}{2} = 3,$$

which are $\{1, 2\}$, $\{1, 3\}$, and $\{2, 3\}$.

Since the order of mutation does not matter, the number of pathways is:

$$\text{number of pathways} = \text{number of genotypes} \cdot 2! = 3 \cdot 2 = 6.$$

The expected waiting time of 2 out of 3 independent mutations is:

$$E(\tau_2) = \sum_P \text{Prob}(P) E(\tau_P),$$

where we have

$$\text{Prob}(P) = \prod_{j=1}^{k} \frac{\lambda_{i_j}}{\sum_{i \in \text{Exit}_i} \lambda_i}, \qquad E(\tau_P) = \sum_{j=1}^{k} \frac{1}{\sum_{i \in \text{Exit}_i} \lambda_i}.$$

For the ease of typing, λ_1, λ_2, and λ_3 will be denoted as A, B, C respectively. Then we have:

$$\begin{aligned}
E(\tau_2) &= \sum_P \text{Prob}(P) E(\tau_P) \\
&= \frac{AB}{(A+B+C)(B+C)} \left(\frac{1}{A+B+C} + \frac{1}{B+C} \right) \\
&+ \frac{AC}{(A+B+C)(B+C)} \left(\frac{1}{A+B+C} + \frac{1}{B+C} \right) \\
&+ \frac{BA}{(A+B+C)(A+C)} \left(\frac{1}{A+B+C} + \frac{1}{A+C} \right) \\
&+ \frac{BC}{(A+B+C)(A+C)} \left(\frac{1}{A+B+C} + \frac{1}{A+C} \right) \\
&+ \frac{CA}{(A+B+C)(A+B)} \left(\frac{1}{A+B+C} + \frac{1}{A+B} \right) \\
&+ \frac{CB}{(A+B+C)(A+B)} \left(\frac{1}{A+B+C} + \frac{1}{A+B} \right) \\
&= \frac{A\cancel{(B+C)}}{(A+B+C)\cancel{(B+C)}} \left(\frac{1}{A+B+C} + \frac{1}{B+C} \right) \\
&\frac{B\cancel{(A+C)}}{(A+B+C)\cancel{(A+C)}} \left(\frac{1}{A+B+C} + \frac{1}{A+C} \right) \\
&+ \frac{C\cancel{(A+B)}}{(A+B+C)\cancel{(A+B)}} \left(\frac{1}{A+B+C} + \frac{1}{A+B} \right) \\
&\text{(Rewrite A, B, C back to } \lambda_1, \lambda_2, \lambda_3 \text{...)} \\
&= \frac{\lambda_1}{\lambda_1+\lambda_2+\lambda_3} \left(\frac{1}{\lambda_1+\lambda_2+\lambda_3} + \frac{1}{\lambda_2+\lambda_3} \right) \\
&+ \frac{\lambda_2}{\lambda_1+\lambda_2+\lambda_3} \left(\frac{1}{\lambda_1+\lambda_2+\lambda_3} + \frac{1}{\lambda_1+\lambda_3} \right) \\
&+ \frac{\lambda_3}{\lambda_1+\lambda_2+\lambda_3} \left(\frac{1}{\lambda_1+\lambda_2+\lambda_3} + \frac{1}{\lambda_1+\lambda_2} \right)
\end{aligned}$$

(c) Now consider d independent mutations. How many paths exist leading to the genotype $\{1, \dots, d\}$ with all mutations present? If cancer already arises after any k mutations, how many different paths are there?

Since all mutations are independent, the number of paths lead to the genotype $\{1, \dots, d\}$ is $d!$.

If cancer already arises after k mutations, then we need to:

1. Choose k mutation out of d
2. Permute the k mutations

to calculate the number of different paths. Therefore, the number of different paths is:

$$\binom{d}{k} \cdot k! = \frac{d!}{k!(d-k)!} \cdot k! = \frac{d!}{(d-k)!}.$$

Problem 2: Neutral Wright-Fisher process

Consider the neutral Wright-Fisher process for a system of N cells of two different types $\{A, B\}$. Let $X(t)$ denote the number of A-cells at time t. The process has the transition matrix

$$P_{i,j} = \text{Prob}[X(t) = j | X(t-1) = i] = \binom{N}{j} \left(\frac{i}{N}\right)^j \left(1 - \frac{i}{N}\right)^{N-j}$$

that is, $X(t)|X(t-1) = i$ is binomially distributed with parameter $p = i/N$.

(a) Compute the conditional expectation $E[X(t)|X(0) = i]$.

We know that

$$X(t) = j | X(t-1) = i \sim \text{Binomial}(N, i/N)$$

so we have

$$E[X(t)|X(t-1) = i] = N \cdot \frac{i}{N} = i = X(t-1).$$

i.e. $E[X(t)|X(t-1)] = X(t-1)$.

By the law of total expectation, we have:

$$\begin{aligned} E_{X(t)}[X(t)|X(0) = i] &= E_{X(t-1)}[E_{X(t)}[X(t)|X(t-1)]|X(0) = i] \\ &= E_{X(t-1)}[X(t-1)|X(0) = i] \\ &= ... \\ &= E_{X(1)}[X(1)|X(0) = i] \\ &= i. \end{aligned}$$

(b) Compute the conditional variance $Var[X(t)|X(0) = i]$.

Hint: Show that

$$Var[X(t)|X(0) = i] = V_1 + \left(1 - \frac{1}{N}\right) Var[X(t-1)|X(0) = i],$$

where $V_1 = Var[X(1)|X(0) = i]$. You can then use the expression above to derive the final result (no explicit calculation needed for this last step).

Similarly, since we have

$$X(t) = j | X(t-1) = i \sim \text{Binomial}(N, i/N),$$

then we have

$$Var[X(t)|X(t-1) = i] = N \cdot \frac{i}{N} \cdot \left(1 - \frac{i}{N}\right) = i\left(1 - \frac{i}{N}\right).$$

Since the mean of the Wight-Fisher process is stationary (as shown in (**a**)), we have $\forall t > 0$, $Var[X(t)|X(t-1) = i] = Var[X(1)|X(0) = 1] = V_1$.

By the law of total variance, we have (similar to homework 2, all the E and Var on the right hand side of the equation are subscripted by $X(t-1)$. The author omitted them out of laziness):

$$\begin{aligned}
Var_{X(t)}[X(t)] &= E[Var[X(t)|X(t-1)]] + Var[E[X(t)|X(t-1)]] \\
&= E\left[X(t-1)\left(1-\frac{X(t-1)}{N}\right)\right] + Var[X(t-1)] \\
&= E[X(t-1)] - \frac{E[X(t-1)^2]}{N} + Var[X(t-1)] \\
&= E[X(t-1)] - \frac{E[X(t-1)]^2 + Var[X(t-1)]}{N} + Var[X(t-1)] \quad (E[A^2] = Var[A] + E[A]^2) \\
&= E[X(t-1)]\left(1-\frac{E[X(t-1))}{N}\right) + \left(1-\frac{1}{N}\right) Var[X(t-1)]
\end{aligned}$$

Applying the fact that $E[X(t)|X(0)=i] = i \forall t$, we have:

$$Var_{X(t)}[X(t)|X(0)=i] = i \cdot \left(1-\frac{i}{N}\right) + \left(1-\frac{1}{N}\right) Var[X(t-1)|X(0)=i]$$

Recall that

$$V_1 = Var[X(1)|X(0)=i] = i\left(1-\frac{i}{N}\right),$$

we can rewrite the equation above as:

$$Var[X(t)|X(0)=i] = V_1 + \left(1-\frac{1}{N}\right) Var[X(t-1)|X(0)=i].$$

(c) Derive an approximation for $Var[X(t)|X(0)=i]$ for large population size N. Compare the variance of the Wright Fisher process to the variance of the Moran process, explain the difference(s).

Let's denote $Var[X(t)|X(0)=i]$ as V_t for simplicity. Then we have:

$$V_t = V_1 + \left(1-\frac{1}{N}\right) V_{t-1}.$$

We can rewrite it into:

$$V_t - a = b(V_{t-1} - a),$$

where

$$\begin{cases} a = \dfrac{V_1}{1/N} \\ b = 1 - \dfrac{1}{N}. \end{cases}$$

From this it follows that (denoting $1/N$ as M for simplicity):

$$
\begin{aligned}
V_t - a &= b^{t-1}(V_{t-1} - a) \\
V_t - \frac{V_1}{M} &= (1-M)^{t-1}\left(V_1 - \frac{V_1}{M}\right) \\
V_t &= V_1(1-M)^{t-1}\frac{M-1}{M} + V_1\frac{1}{M} \\
V_t &= V_1\frac{1-(1-M)^t}{M}
\end{aligned}
$$

For large population size $N \to \infty$, we have $M = 1/N \to 0$, then we have:

$$
\begin{aligned}
V_t &= V_1 \cdot \lim_{M\to 0} \frac{1-(1-M)^t}{M} \\
&= V_1 \cdot \lim_{M\to 0} \frac{1-(1-tM+O(M^2))}{M} \qquad \text{(Taylor expansion)} \\
&= V_1 \cdot t.
\end{aligned}
$$

Rewrite it in the desired form:

$$
Var[X(t)|X(0)=i] = V_1 \cdot t.
$$

Recall from homework 2 (and thanks for all the hints in homework 2 that made the author possible to solve homework 5), the Moran variance of Moran process is (note that now $V_1 = 2(i/N)(1-i/N)$ for Moran process):

$$
Var[X(t)|X(0)=i] = V_1 \cdot \frac{1-(1-2/N^2)^t}{2/N^2},
$$

which has very similar form to that of the Wright-Fisher process, and for large population size N, we also have $Var[X(t)|X(0)=i] = V_1 \cdot t$.

The difference between the two processes is that the Moran process has a factor of $2/N^2$ in the denominator, which makes the variance of the Moran process smaller than that of the Wright-Fisher process. This is natural because one generation in Wright-Fisher process is N times larger than that in the Moran process. The variance is not exactly N times larger for Wright-Fisher process, though, since the Moran process is modelling more than merely genetic drift but also selection and competition.

(d) See below.

Show that in the Wright-Fisher process, the heterozygosity H_t at time t satisfies

$$
E[H_t|X_0 = i] = H_0(i)\left(1-\frac{1}{N}\right)^t,
$$

and hence decreases exponentially at rate $1/N$. Compare this behaviour with the Moran model.
Note: Heterozygosity in this context is defined as the probability that two individuals chosen at random from the population are of different types.

By definition,

$$\begin{aligned} H_t &= 2 \cdot \frac{X(t)}{N} \cdot \left(1 - \frac{X(t)}{N}\right) \\ &= 2 \cdot \left(\frac{X(t)}{N} - \frac{X(t)^2}{N^2}\right) \end{aligned}$$

With $X_0 = i$, $H_0 = 2 \cdot i/N \cdot (1 - i/N)$.

$$\begin{aligned} E[H_t] &= E\left[2 \cdot \left(\frac{X(t)}{N} - \frac{X(t)^2}{N^2}\right)\right] \\ &= 2 \cdot \left(\frac{E[X(t)]}{N} - \frac{E[X(t)^2]}{N^2}\right) \\ &= 2 \cdot \left(\frac{E[X(t)]}{N} - \frac{Var[X(t)] + E[X(t)]^2}{N^2}\right) \end{aligned}$$

Based on previous knowledge, we have

$$E[X(t)|X_0 = i] = i, \qquad Var[X(t)|X_0 = i] = V_1 \cdot \frac{1 - (1 - 1/N)^t}{1/N},$$

where $V_1 = N \cdot i/N \cdot (1 - i/N)$.

Then we have:

$$\begin{aligned} E[H_t|X_0 = i] &= 2 \cdot \left(\frac{i}{N} - \frac{i^2}{N^2} - \frac{V_1 \cdot \frac{1-(1-1/N)^t}{1/N}}{N^2}\right) \\ &\text{(Denote } i/N \text{ as } p \text{ for simplicity)} \\ &= 2 \cdot \left(p(1-p) - \frac{N^2 p(1-p) \cdot (1 - (1 - 1/N)^t)}{N^2}\right) \\ &= 2\left(p(1-p) - p(1-p) + p(1-p)(1-1/N)^t\right) \\ &= 2p(1-p) \cdot (1 - 1/N)^t. \end{aligned}$$

As by definition,

$$H_0|_{X_0=i} = 2 \cdot \frac{i}{N} \cdot \left(1 - \frac{i}{N}\right) = 2p(1-p).$$

We can rewrite the above equation into the desired form:

$$E[H_t|X_0 = i] = H_0|_{X_0=i} \cdot (1 - 1/N)^t. = H_0(i) \cdot (1 - 1/N)^t.$$

For the Moran process, the derivation process is similar, we just substitute $1/N$ to $2/N^2$ and $V_1 = 2(i/N)(1 - i/N) = 2p(1-p)$

$$\begin{aligned} E[H_t|X_0 = i] &= 2 \cdot \left(p(1-p) - \frac{2p(1-p) \cdot \frac{1-(1-2/N^2)^t}{2/N^2}}{N^2}\right) \\ &= 2 \cdot \left(p(1-p) - p(1-p) \cdot (1 - (1 - 2/N^2)^t)\right) \\ &= 2p(1-p)(1 - 2/N^2)^t \end{aligned}$$

Similarly, we also have $H_0|_{X_0=i} = 2p(1-p)$ for Moran process, then we have:

$$E[H_t|X_0 = i] = H_0(i) \cdot (1 - 2/N^2)^t,$$

which is extremely similar to that of the Wright-Fisher process.

Problem 3: Wave approximation

> Consider the wave approximation of the Wright-Fisher model for cancer progression. Here, the growth of a clone with j mutations is given by
>
> $$\dot{x}_j = s x_j (j - \langle j \rangle).$$
>
> For small times, the average fitness $s\langle j \rangle = s \sum_j x_j$ can be considered constant. Use this throughout your calculations.

(a) Find the analytic solution for the initial condition $x_j(0) = 1/N$.

Since the average fitness $s\langle j \rangle$ is constant, the above equation is equivalent to a simple first-order linear homogeneous ODE:

$$\dot{x}_j = \alpha \cdot x_j,$$

where

$$\alpha = s(j - \langle j \rangle).$$

The solution to that ODE is

$$x_j(t) = A_0 e^{\alpha t}.$$

Substitute the initial condition $x_j(0) = 1/N$ into the equation above, we have:

$$x_j(t) = \frac{1}{N} \exp\left(s(j - \langle j \rangle)t\right).$$

(b) The rate at which an additional mutation in a clone with j mutations occurs is given by $udx_j(t)$. Find the time τ when the cumulative probability exceeds $1/N$.

The rate at which an additional mutation in a clone with j mutations occurs is given by $udx_j(t)$, which means at time t there is expected to be $udx_j(t)$ event. The cumulative probability of the event is:

$$P = \int_{t=0}^{\tau} udx_j(t)$$

When it exceeds $1/N$, we have (denote $s(j - \langle j \rangle)$ as A for simplicity):

$$
\begin{aligned}
\frac{1}{N} &= \int_{t=0}^{\tau} u d x_j(t) dt. \\
\frac{1}{N} &= \int_{t=0}^{\tau} \frac{1}{N} u d \exp(At) dt \\
1 &= ud \int_{t=0}^{\tau} \exp(At) dt \\
1 &= \frac{ud}{A} \left[\exp(At)\right]_{t=0}^{\tau} \\
1 &= \frac{ud}{A} \left(\exp(A\tau) - 1\right) \\
\frac{A}{ud} + 1 &= \exp(A\tau) \\
A\tau &= \ln\left(\frac{A + ud}{ud}\right) \\
\tau &= \frac{\ln\left(\frac{A+ud}{ud}\right)}{A}
\end{aligned}
$$

To rewrite it in the desired way:

$$
\tau = \frac{\ln\left(\frac{s(j-\langle j\rangle)+ud}{ud}\right)}{s(j - \langle j\rangle)}
$$

(c) Compute the waiting time until the next mutation.

> Question settings: mutation rate $u = 10^{-7}$ per cell generation, $d = 80$ genes and a fitness advantage of $s = 1.15\%$ per mutation. Use that $j - \langle j\rangle \approx \sqrt{\ln N}$ with $N = 10^7$ cells and assume a cell generation time of 1 day.

From the slides we know that the average time it takes until the next mutation appears is:

$$
\tau = \frac{\ln[s/ud]^2}{2s \ln N}
$$

From the equation in (**b**), we have:

$$
\tau = \frac{\ln\left(\frac{s(j-\langle j\rangle)+ud}{ud}\right)}{s(j - \langle j\rangle)} \approx \frac{\ln\left(\frac{s\sqrt{\ln N}+ud}{ud}\right)}{s\sqrt{\ln N}}
$$

The author didn't really know what to calculate so chose to do both.

```
u <- 10^(-7)
d <- 80
s <- 0.0115
N <- 10^7
```

```
tau_slides <- log(s / u * d)^2 / (2 * s * log(N))
```

The waiting time until the next mutation is 693.55 days.

```
A <- s * sqrt(log(N))
tau_b <- log((A + u*d)/u*d) / A
```

The waiting time until the next mutation is 377.41 days.

(Feel free to give deductions and comments as I really didn't understand this part.)