hapter 1 What is Evolution?

Definition of Evolution

volution is the change in the frequency of different types of individuals in population over time.

Biological evolution is the change in allele frequencies within a gene pool

Evolution generates inheritable traits and is driven by replication, mutation, selection, dispersal, environment, interactions

Exponential Growth (unrealistic model in the long run)

Discrete case :

x_t := number of cells in generation t

$$\underset{\text{doubling}}{X_{t+1} = 2X_t} \Rightarrow X_t = X_0 2^t$$

Continuous case :

$x(t)$:= continuous number / fractions of cells at time t

T := generation time / time from birth to reproduction

$$T \sim \exp\left(\frac{1}{r}\right), \; r := \text{growth rate}$$

$$\Rightarrow \mathbb{P}(T \leq t) = 1 - e^{-rt}$$

$$\frac{dX(t)}{dt} = rX(t) \Rightarrow X(t) = X_0 e^{rt}$$

- Cell Death

d := death rate $\Rightarrow$ $\frac{1}{d}$:= average life span

$\Rightarrow$ modified differential equation :

$$\frac{dX(t)}{dt} = (r-d)X(t) \Rightarrow X(t) = X_0 e^{(r-d)t}$$

- Basic Reproductive Ratio : expected number of offspring of an individual

$$R_0 = \frac{r}{d}$$

↳ $R_0 > 1$: population expands indefinitely

↳ $R_0 < 1$: population goes extinct

↳ $R_0 = 1$: population size remains constant

but $x^* = x_0$ is not stable

(4) Logistic Growth

$K :=$ Carrying capacity of the population

- Logistic Equation :

$$\frac{dX(t)}{dt} = rX(t)\left(1 - \frac{X(t)}{K}\right)$$

$$\Rightarrow X(t) = \frac{K X_0 e^{rt}}{K + X_0(e^{rt} - 1)}$$

- Equilibrium Points :

(1) $x_1^* = 0$; stable if $r<0$ and not stable if $r>0$

(2) $x_2^* = K$: stable if $r>0$ and not stable if $r<0$

(5) Stability of Equilibrium Points

- Discrete case :

$$x_{t+1} = f(x_t) \Rightarrow x^* = f(x^*)$$

$\Rightarrow x^*$ is attractive if $|f'(x^*)| < 1$

x^* is repelling if $|f'(x^*)| > 1$

- Continuous case :

$$\frac{dx(t)}{dt} = f(x(t)) \Rightarrow f(x^*) = 0$$

$\Rightarrow x^*$ is attractive if $f'(x^*) < 0$

x^* is repelling if $f'(x^*) > 0$

(6) Logistic Difference Equation

$$x_{t+1} = rx_t(1 - x_t)$$, $x_t :=$ fractions of cells

- The number of equilibrium points x^* depends on the value of r
  - ↳ If $r<1$, then the point $x^* = 0$ is stable $\Rightarrow$ population goes extinct
  - ↳ If $r>4$, then the system will diverge to $-\infty$ $\Rightarrow$ population goes extinct
  - ↳ For r moving from 1 to 4, the number of equilibrium points grows exponentially and eventually becomes chaotic. (around $r=3.57$)

(I) Selection

(1) 2 independent exponentially growing types

Consider A : growth rate $a > 0$, abundance $x(t)$ $\Rightarrow$ $\frac{dx(t)}{dt} = ax(t)$

B : growth rate $b > 0$, abundance $y(t)$ $\Rightarrow$ $\frac{dy(t)}{dt} = by(t)$

Relative abundance

$$\rho(t) = \frac{x(t)}{y(t)}$$

$$\Rightarrow \quad \frac{d\rho(t)}{dt} = \frac{x'(t)\,y(t) - y'(t)\,x(t)}{y^2(t)} = (a-b)\,\rho(t)$$

$$\Rightarrow \quad \rho(t) = \rho_0\, e^{(a-b)t}$$

↳ If $a > b$, then $\rho \to \infty$ and selection favors A over B (not extinct)

↳ If $a < b$, then $\rho \to 0$ and selection favors B over A (not extinct)

↳ If $a = b$, then $\rho = \rho_0$.

(2) 2 Competing types

$x(t) :=$ relative abundance of A , $y(t) :=$ relative abundance of B

Constraint :

$$x(t) + y(t) = 1 \quad \forall\ t > 0 \qquad (*)$$

Dynamical system :

$$x'(t) = x(t)\,(a - \phi)$$
$$y'(t) = y(t)\,(b - \phi)$$

where $\phi := ax(t) + by(t)$: average fitness of the population

↳ ensures that constraint $(*)$ is satisfied

By substituting $y(t) = 1 - x(t)$, we get

$$\frac{dx(t)}{dt} = (a-b)\,x(t)\,(1 - x(t))$$

Equilibrium points :

(1) $x^* = 1$: all-A state , stable if $a > b$

(2) $x^* = 0$: all-B state , stable if $a < b$

(3) n Competing Types

- Probability simplex :

$$S_n = \{ (x_1, \dots, x_n) \mid x_i \geq 0 , \sum_{i=1}^{n} x_i = 1 \}$$

- Type i, $i \in \{1, \dots, n\}$: fitness f_i, frequency $x_i(t)$
- The type frequencies are points in the $(n-1)$-dimensional probability simplex S_n

$$x_1(t) + \cdots + x_n(t) = 1$$

- Dynamical system :

$$x_i'(t) = x_i(t) (f_i - \phi(t)) , \quad i = 1, \dots, n$$

where $\phi(t) = x_1(t) f_1 + \cdots + x_n(t) f_n$: average fitness of the population

- Single Equilibrium Point : starting from any interior point in S_n, the fittest type will eventually outcompete all others (survival of fittest)

(4) Subexponential and superexponential growth

$$x'(t) = a x^c - \phi x$$
$$y'(t) = b y^c - \phi y$$

where $x(t) + y(t) = 1 \quad \forall t \geq 0$, $\Phi(t) = a x(t)^c + b y(t)^c$

$$\Leftrightarrow \quad x'(t) = x(t) (1 - x(t)) f(x(t))$$

where $f(x(t)) = a x^{c-1} - b(1-x)^{c-1}$

- $c = 0$: growth is linear (immigration)
- $c = 1$: exponential growth
- $c > 1$: superexponential growth
- $c < 1$: subexponential growth

Fixed points :

1. $x = 0$
2. $x = 1$
3. For $c \neq 1$, $x^* = \dfrac{1}{1 + \left(\frac{a}{b}\right)^{\frac{1}{c-1}}} \in (0, 1)$ (mixed population)

urvival of all ↳ $c < 1$: x^* is globally stable, whereas all-A and all-B are not (less fit type can invade)

urvival of first ↳ $c > 1$: x^* is unstable ($x > x^*$: all-A ; $x < x^*$: all-B)
(less fit can prevent invasion)

(8) Mutation

- Mutation can occur with/without reproduction.

(1) Basic Mutation Dynamics :

Assume fitness $a = b = 1$. Let mutation rates $u_1 = P(A \to B)$ and $u_2 = P(B \to A)$

$$x' = x(1-u_1) + yu_2 - \phi x$$
$$y' = y(1-u_2) + xu_1 - \phi y$$

where $\phi = ax + by = 1$ and $x + y = 1$

$$\Leftrightarrow \quad x' = u_2 - x(u_1 + u_2)$$
$$y' = u_1 - y(u_1 + u_2)$$

$$\Leftrightarrow \quad x^* = \frac{u_2}{u_1 + u_2}, \quad y^* = \frac{u_1}{u_1 + u_2} \quad \Rightarrow \quad \frac{x^*}{y^*} = \frac{u_2}{u_1}$$

$\Rightarrow$ Mutation can alone lead to coexistence if u_1 and u_2 are similar

$\Rightarrow$ if $u_1 >> u_2$, we may assume $u_2 = 0$ and A can go extinct.

Mutation can alone lead to extinction.

(2) Mutation Dynamics of n types

Let Q be an $n \times n$ matrix with entries

$$q_{ij} = P(\text{type } i \Rightarrow \text{type } j), \quad \sum_{j=1}^{n} q_{ij} = 1$$

$\Rightarrow$ Q is a stochastic matrix

Dynamical system :

$$x_i' = \sum_{j=1}^{n} x_j q_{ji} - \phi x_i, \quad i = 1, \dots, n$$

$$\Leftrightarrow \quad x' = xQ - \phi x$$

where $\phi = x_1 f_1 + \cdots + x_n f_n$: average fitness of the population

Equilibrium point :

$x^* :=$ left eigenvector of Q associated with the largest eigenvalue 1.

↳ asymmetric mutation can result in selection even without growth differences.

Chapter 2 Quasispecies

1) Central Dogma of Molecular Biology

$$\text{DNA} \xrightarrow{\text{transcribes}} \text{RNA} \xrightarrow{\text{translates}} \text{Protein}$$

2) RNA Virus

- use the transcription machinery of the host cell to replicate
- short genome
- very high mutation rate (no proof-reading of reverse transcription)
- high turnover
- exposed to strong selective forces
- extreme evolutionary dynamics

3) Sequence Space

- Sequence of length L :

$$A^L = \{(a_1, \dots, a_L) : a_i \in \underset{\text{alphabet}}{\underset{\downarrow}{A}}\}$$

- Sequence space : $A^* = \bigcup_{L \geq 0} A^L$
- # of sequences in $A^L := |A|^L$
- Hamming distance :

$$\|x - y\| = \sum_{i=1}^{L} \mathbb{1}\{x_i \neq y_i\} \leq L$$

- Evolution is a trajectory of a population in sequence space.
- Genotype space :

$$G \subseteq A^*$$

- Fitness Landscape :

$$f : G \to \mathbb{R} \quad \text{s.t.} \quad f(g) \in \mathbb{R}$$

↳ for individual : discrete ; for average population fitness : continuous

↳ epistasis : $\varepsilon = (f_{00} + f_{11}) - (f_{01} + f_{10})$ ⇒ measure the strength of additive effects

④ The Quasispecies Equation

Let $x(t) = (x_0(t), \dots, x_n(t))$ be the genotype frequencies for $i = 1, \dots, n$ genotypes at t

Let $Q = (q_{ij}) = (q_{i \to j})$ be a mutation matrix.

Let $f = (f_0, \dots, f_n)$ be a fitness landscape.

Denote by $\phi = x^T f$ the average fitness of the population.

$$\dot{x}_i = \frac{dx_i}{dt} = \sum_{j=0}^{n} x_j \underset{\substack{\downarrow \\ \text{selection}}}{f_j} \underset{\substack{\downarrow \\ \text{mutation}}}{q_{ji}} - \phi x_i \quad , \; i = 1, \dots, n$$

- No mutation: $Q = I$

  ↳ we recover the selection equation ("survival of the fittest")
- No selection: $f = (1, \dots, 1)$

  ↳ we recover the mutation equation
- If Q is irreducible, there exists a globally stable equilibrium x^* inside S_{n+1} but x^* does NOT maximize the fitness ϕ.
- Mutation-Selection Matrix:

$$W = (w_{ij}) = (f_j q_{ji})$$

$$\Rightarrow \quad \dot{x} = Wx - \phi x$$

↳ if W is non-negative and irreducible, then the largest eigenvalue of W corresponds to the average fitness ϕ, and the associated eigenvector after normalization is the global equilibrium point. = mutation-selection balance

⑤ Adaptation (HIV wants to be close to the error threshold but not over it)

- adaptation of a population is localization in sequence space at a local maximum of the fitness landscape
- selection drives the population towards the peak, whereas mutation drives away
- error threshdel: necessary condition on mutation rate for adaptation to occur

  ↳ a simplified case: x_0 wild type with fitness $f_0 > 1$, x_1 other types with fitness 1

  Let $u :=$ mutation rate $\Rightarrow$ wild type is copied error-free w̄ $q = (1-u)^L$

  Ignoring back mutation, we get

$$\begin{cases} \dot{x}_0 = x_0 f_0 q - \phi x_0 \\ \dot{x}_1 = x_0 f_0 (1-q) + x_1 - \phi x_1 \end{cases}$$

$$\Rightarrow \quad x_0^* = \begin{cases} \dfrac{f_0 q - 1}{f_0 - 1} & \text{if } f_0 q > 1 \\ 0 & \text{otherwise} \end{cases}$$

If $u \ll 1$ and $\log f_0 \approx 1$, then

$\log(f_0 q) = 1 + L \log(1-u) \approx 1 - Lu > \log 1 =$

$\Leftrightarrow uL < 1$ (expected mutation per replicatio

$\Leftrightarrow u_c := \frac{1}{L}$

If $uL > 1$, have mutational meltdown

Chapter 3 : Stochastic Models of Finite Populations

1) Probability

- Exponential distribution : $X \sim \exp(\lambda)$, $f(x) = \lambda e^{-\lambda x}$, $x \geq 0$

  $P(X \leq x) = 1 - e^{-\lambda x}$, $P(X > x) = e^{-\lambda x}$

  $E[X] = \frac{1}{\lambda}$, $Var[X] = \frac{1}{\lambda^2}$

  ↳ memoryless property : $P(X > s+t) = P(X > s)\,P(X > t)$

  $P(X > s+t \mid X > t) = P(X > s)$

  ↳ Competing exponentials :

  $X \sim \exp(\lambda)$, $Y \sim \exp(\mu)$, and $X \perp\!\!\!\perp Y$

  $\Rightarrow$ $\min(X, Y) \sim \exp(\lambda + \mu)$ and $P(X < Y) = \frac{\lambda}{\lambda + \mu}$

- Markov chain :

  $$\{X(t) \mid t \in T = \{0, 1, 2, \ldots\}\}$$

  with $P(X(t+1) \mid X(0), \ldots, X(t)) = P(X(t+1) \mid X(t))$

  ↳ transition matrix : $P = (P_{ij})$, $P_{ij} = P(X(t+1) = j \mid X(t) = i)$

  ↳ time-homogeneous Markov chain : $P(X(t+1) = j \mid X(t) = i) = P(X(1) = j \mid X(0) = i)$

  ↳ absorbing state : $X(t) = x^* \quad \forall\, t \geq t_0$

  ↳ An ergodic Markov chain has a unique stationary distribution π such that

  $$\pi^T P = \pi^T$$

  where $\lim_{t \to \infty} P_{ij}(t) = \pi_j$, $\forall\, i, j \in S$:= state space

2) The Moran Process

(1) Model :

- Population size : $N < \infty$.
- 2 types of individuals A and B
- process :

  (1) pick randomly an individual for reproduction

  (2) pick randomly an individual for death

  (3) the offspring of the first individual replaces the second individual

- Both types have the same probability of reproduction and death.

  $\Rightarrow$ Neutral drift : changes in allele frequency are only due to random fluctuations.

(2) Birth-Death Process

The state space is $\{0, 1, \dots, N\}$. Let $i :=$ # of A individuals.

Let $p = \frac{i}{N}$:= allele frequency of A.

Then, the transition matrix is given by

$$
\begin{aligned}
P_{i,i+1} &= p(1-p) \\
P_{i,i-1} &= (1-p)p \\
P_{i,i} &= 1 - 2p(1-p) = p^2 + (1-p)^2
\end{aligned}
$$

↳ P is a tri-diagonal matrix with all other entries zero

↳ it is a birth-death process, because the # of A individuals can only change one step at a time

(3) Absorbing States

1. All-A individuals : $P_{N,N} = 1$ and $P_{N,i} = 0 \ \forall i < N$
2. All-B individuals : $P_{0,0} = 1$ and $P_{0,i} = 0 \ \forall i > 0$

(4) Fixation Probabilities :

Let x_i := probability of ending up in state N when starting from state i

Then, in the Moran (neutral) process, we must have

$$x_i = \frac{i}{N} \quad , \quad i = 0, \dots, N$$

because each cell has the same chance to produce offspring, then each cell also has the same chance to be the origin of cells eventually in the population, and thus the same chance of fixation.

(5) Stationary Mean :

$$\mathbb{E}[X(t) \mid X(0) = i] = i$$

(6) Time-dependent Variance :

$$\lim_{N \to \infty} Var(X(t) \mid X(0) = i) = V_1 t$$

where $V_1 = Var(X(1) \mid X(0) = i) = 2p(1-p) = 2\frac{i}{N}(1 - \frac{i}{N})$

(7) General Birth-Death Process :

$$P_{i,i+1} = \alpha_i , \quad P_{i,i-1} = \beta_i \quad \Rightarrow \quad \gamma_i = \frac{\beta_i}{\alpha_i} , \quad \alpha_0 = \beta_N = 0$$

$$\Rightarrow \quad x_i = \frac{1 + \sum_{j=1}^{i-1} \prod_{k=1}^{j} \gamma_k}{1 + \sum_{j=1}^{N-1} \prod_{k=1}^{j} \gamma_k} \quad , \quad i = 0, \dots, N$$

8) Mean Fixation Time

$$-N^2[(1-p)\log(1-p) + p\log p]$$

9) Heterozygosity

$H_t :=$ probability that two individuals chosen at random from the population are of different types.

$$\mathbb{E}[H_t \mid X(0)=i] = H_0(i)\left(1-\frac{2}{N^2}\right)^t, \quad H_0(i) = 2\cdot\frac{i}{N}\cdot\frac{N-i}{N-1}$$

↳ decays exponentially at rate $\frac{2}{N^2}$

↳ quantifies the amount of random drift that the population is experiencing

3) Moran Process with Constant Selection

Consider 2 types of individuals A and B with growth rates

$$\lambda_A = r \quad \text{and} \quad \lambda_B = 1$$

Let the waiting time to the next birth of A be

$$T_A \sim \min\{\underbrace{\exp(\lambda_A), \dots, \exp(\lambda_A)}_{i \text{ individuals}}\} = \exp(ir)$$

Similarly,

$$T_B \sim \exp((N-i)\lambda_B) = \exp(N-i)$$

Then,

$$P(T_A < T_B) = \frac{ir}{ir+N-i}, \quad P(T_B < T_A) = \frac{N-i}{ir+N-i}$$

The transition probabilities become

$$P_{i,i+1} = \underbrace{\frac{ri}{ri+N-i}}_{P(T_A<T_B)} \cdot \underbrace{\frac{N-i}{N}}_{\text{choose a } B \text{ to die}}$$

$$P_{i,i-1} = \underbrace{\frac{N-i}{ri+N-i}}_{P(T_A>T_B)} \cdot \underbrace{\frac{i}{N}}_{\text{choose an } A \text{ to die}}$$

$$P_{i,i} = 1 - P_{i,i+1} - P_{i,i-1}$$

With $\gamma_i = P_{i,i-1}/P_{i,i+1} = \frac{1}{r}$ $\forall i = 1,\dots,N-1$, $\gamma_0 = 0$ and $\gamma_N = 1$, the fixation probabilities become

$$x_i = \frac{1-1/r^i}{1-1/r^N}, \quad i = 0,\dots,N$$

In particular,

$$p = x_1, \quad \lim_{r\to 1} p = \frac{1}{N} \quad \text{(neutral Moran process)}$$

④ Poisson Process

- A Poisson process is a stochastic counting process, a continuous-time Markov chain with independent Poisson distributions in each interval.
- Model :

$$\{N(t) \mid t \geq 0\} \quad \text{with} \quad N(0) = 0$$

$$N(t+s) - N(s) \sim \text{Poisson}(\lambda t) \Rightarrow \mathbb{P}(N(t+s) - N(s) = k) = \frac{e^{-\lambda t}(\lambda t)^k}{k!}$$

- Inter-arrival times are exponential :

$$\{T_n \mid n = 1, 2, \dots\}$$

$$T_n \sim \exp(\lambda), \quad n = 1, 2, \dots$$

sketch proof :

$$\mathbb{P}(T_1 > t) = \mathbb{P}(N(t) = 0) = e^{-\lambda t}$$

$$\mathbb{P}(T_2 > t) = \mathbb{E}_{T_1}[\mathbb{P}(T_2 > t) \mid T_1] \quad \text{by law of total expectation}$$

$$= \int_s \mathbb{P}(N(s+t) = N(s) \mid T_1 = s)\, f_{T_1}(s)\, ds$$

$$= \int_s \mathbb{P}(N(t) = 0)\, f_{T_1}(s)\, ds$$

$$= e^{-\lambda t}$$

⇒ proof by induction

⑤ Rate of Evolution

Consider an all-A population where a B mutant occur at mutation rate $u << 1$. Then the number of mutations over time can be modeled by the Poisson process. The waiting time for the first mutant to occur is

$$T_1 \sim \exp(\underbrace{Nu}_{\text{total mutation rate}})$$

Suppose B has an selective advantage r and the fixation probability is $\rho = x_1$.

⇒ Rate of Evolution from all-A to all-B :

$$R = \underbrace{Nu}_{\text{prob. of first mutant to occur}}\rho \rightarrow \text{prob. of that mutant to fixate}$$

↳ If $r = 1$, then $\rho = \frac{1}{N}$ and $R = u$, which means in the neutral case, the rate of evolution is independent of the population size and depends solely on the mutation rate.

Chapter 4 Evolutionary Dynamics of Cancer

① Cancer

- Cancer is a breakdown of cellular cooperation for multicellular organisms.
- The somatic evolution of cancer is the uncontrolled, selfish replication of cells, which give rise to tumors.
- Cancer progression: accumulation of mutations in genes, which will increase somatic fitness of cells (apoptosis escape)
- Most cancer cells are aneuploid (strange copy numbers)
- normal → adenoma → cancer → metastasis

② Oncogenes

- Oncogenes increase fitness if one allele is mutated or inappropriately expressed.
- Activation: (1) Point Mutation (2) Gene Amplification (3) Chromosomal Fusion
- Fixation: Moran process in a compartment with initially normal cells

  ↳ population size N ↳ mutation rate u ↳ fitness advantage r

  ↳ fixation probability if there is one mutant:

$$\rho = x_1 = \frac{1 - 1/r}{1 - 1/r^N}$$

  ↳ probability that a mutant has been fixed by time t:

$$P(t) = 1 - e^{-\underbrace{Nu\rho}_{\downarrow} t}$$

rate of evolution from all-A to all-B

(1) $r > 1$: advantageous fitness, $P(t)$ increases as N increases

↳ large compartments accelerate accumulation of mutations

↳ to prevent fixation of a mutant, most tissues with high turnover are organized in many small compartments.

(2) $r < 1$: deleterious fitness, $P(t)$ decreases as N increases

(3) $r = 1$: fixation of a mutant (new) is independent of N

↳ $P(t) = 1 - e^{-ut}$

③ Linear Process of Cancer

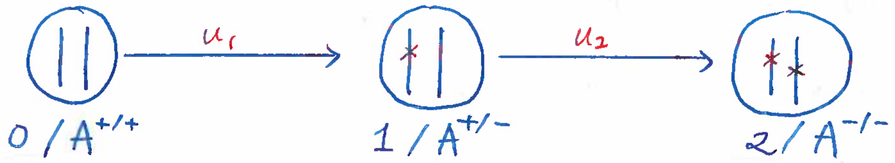
- Architecture: one or few stem cells that perform asymmetric division into stem cells and cells that are more differentiated. The cells are pushed linearly towards the edge and will eventually undergo apoptosis.
- Only mutations on stem cells can lead to fixation, while all other mutations are "washed ou
- Fixation probability is independent of the fitness advantage r: $P = \frac{1}{N}$ (one stem cell
- Perfect design to protect against mutations in tumour suppressor genes/oncogenes but not instabilit

④ Tumour Suppressor Genes (TSG)

- TSGs increase fitness if both alleles are mutated by
  (1) two point mutations
  (2) one point mutation + loss of heterozygosity (LOH)

  ↳ TSGs are inactivated

- Dynamics of TSG Inactivation



Goal: In a population of size N, what is the probability that at least one cell has been hit by 2 mutations by time t? ⇒ $\mathbb{P}(t)$

(1) Small Population size

- Population size: N
- Mutation rates: u_1, u_2

⇒ Fixation probability of the first mutation: (neutral)

$$\rho = x_1 = \frac{1}{N}$$

↳ time until first fixation:

$$T_1 \sim \exp\left(\frac{1}{N}\right)$$

$$\Rightarrow \mathbb{E}[T_1] = N = \text{population size}$$

⇒ Waiting time for the second mutation to occur in any cell:

$$T_2 := \min\{\tilde{T}_1, \dots, \tilde{T}_N\} \sim \exp(Nu_2)$$

$$\Rightarrow \mathbb{E}[T_2] = \frac{1}{Nu_2}$$

⇒ Type 1 cells reach fixation before a type 2 cell arises:

$$N << \frac{1}{Nu_2} \iff N << 1/\sqrt{u_2} \quad \text{(definition of small population size)}$$

- Dynamical System:

  State 0: all type 0 ; State 1: all type 1 ; State 2: 1 type 2

$$\dot{X}_0 = -Nu_1\rho X_0 = -N \cdot u_1 \cdot \frac{1}{N} X_0 = -u_1 X_0$$

$$\dot{X}_1 = u_1 X_0 - Nu_2 X_1$$

$$\dot{X}_2 = Nu_2 X_1$$

where X_0, X_1, X_2 are probabilities

In particular, $P(t) = X_2(t)$

(2) Intermediate Population Size

- Average waiting time for a type 1 cell to occur is

$$\frac{1}{Nu_1}$$

↳ we say that it is long if $\frac{1}{Nu_1} > 1 \iff N < \frac{1}{u_1}$

- A type 2 cell is generated before the fixation of type 1 cell if

$$N > 1/\sqrt{u_2}$$

⇒ Intermediate regime for tunneling :

$$\frac{1}{\sqrt{u_2}} << N << \frac{1}{u_1}$$

s.t.

$$\begin{tikzcd} 0 \arrow[r] \arrow[rr, bend right] & 1 \arrow[r] & 2 \end{tikzcd}$$

(3) Large Population Size :

$$N >> \frac{1}{u_1}$$

↳ type 1 cells will be generated immediately (instantaneous waiting time) and they grow linearly according to

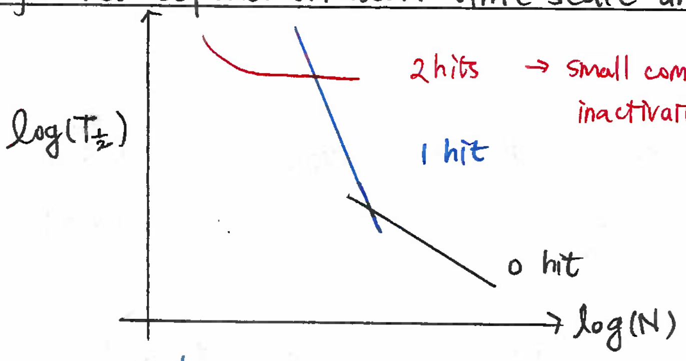
$$x_1(t) = Nu_1 t$$

↳ probability of generating a type 2 cell during type 1 growth is

$$\mathbb{P}(t) = 1 - \exp\Big(-u_2 \underbrace{\int_0^t x_1(t)\,dt}_{\text{abundance}}\Big) \quad \text{of type 1 at time } t$$

$$= 1 - \exp\left(-\tfrac{1}{2} N u_1 u_2 t^2\right)$$

- TSG Inactivation Dynamics depend on both time scale and population size :



→ small compartments protect against the inactivation of tumour suppressor genes

(1) Short time scale : $t << \frac{1}{Nu_2}$, we have $P(t) \approx N u_1 u_2 t^2/2$ (2 rate limiting events

(2) Intermediate time scale : $\frac{1}{Nu_2} << t << \frac{1}{u_1}$, $\mathbb{P}(t) \approx 1 - e^{-u_2 t}$ (first hit is rate limiting

(3) Long time scale : $t >> \frac{1}{u_1}$, $P(t) = 1 - \exp(-\frac{1}{2} N u_1 u_2 t^2)$ (no rate limiting events

(5) Chromosomal Instability (CIN)

- CIN leads to increased rate of gaining or losing whole chromosomes or large parts of it $\overset{(LOH)}{}$ ⇒ inactivation of TSGs ($u \approx 10^{-2}$)

- Model :

$$\begin{tikzcd}
A^{+/+} \arrow[r, "u_1"] \arrow[d, "u_c"'] & A^{+/-} \arrow[r, "u_2"] \arrow[d, "u_c"'] & A^{-/-} \arrow[d, "u_c"'] \\
A^{+/+}_{CIN} \arrow[r, "u_1"] & A^{+/-}_{CIN} \arrow[r, "u_3"] & A^{-/-}_{CIN}
\end{tikzcd}$$

(1) Neutral CIN :

Consider small compartments : $N << \frac{1}{u_1}, \frac{1}{u_2}, \frac{1}{u_c}$

Assume $A^{+/-}$ and CIN cells are neutral and $A^{-/-}$ will be fixed immediately

$$\begin{tikzcd}
X_0 \arrow[r, "u_1"] \arrow[d, "u_c"'] & X_1 \arrow[r, "Nu_2"] \arrow[d, "u_c"'] & X_2 \\
Y_0 \arrow[r, "u_1"] & Y_1 \arrow[r, "Nu_3", "(fast)"'] & Y_2
\end{tikzcd}$$

⇒ $X_2(t) \approx N u_1 u_2 t^2/2$ and $Y_2(t) \approx Y_1(t) \approx u_1 u_c t^2$

⇒ waiting time for LOH is negligible since $u_3 \approx 10^{-2}$!

(2) Costly CIN in small compartment :

Assume CIN cells have fitness advantage $r < 1$.

⇒ fixation probability of a CIN cell :

$$\rho = \frac{1 - 1/r}{1 - 1/r^N}$$

⇒ non-CIN cell state to all-CIN cell state rate of evolution :

$$N u_c \rho$$

⇒ $X_2(t) \approx N u_1 u_2 t^2/2$ and $Y_2(t) \approx Y_1(t) \approx N\rho u_1 u_c t^2 < u_1 u_c t^2$

(3) Costly CIN in large compartment :

For large N, $N\rho$ becomes vanishingly small, so intermediate CIN types will not fixate. The population tunnels from X_1 to Y_2 at rate

$$R = \frac{N u_c u_3 r}{1 - r}$$

⇒ $X_2(t) \approx N u_1 u_2 t^2/2$ and $Y_2(t) \approx R u_1 t^2/2$, $Y_0(t) = Y_1(t) = 0$

⇒ Large compartment protects against the fixation of CIN cells

⇒ Linear design (eg. colon crypts) is the best design to protect against oncogenes and TSGs but not to genetic instability

Chapter 5 Cancer Progression

1) Waiting Times

(1) Multistage Theory

Assume that tumour progression follows a linear, multistep process where each step has mutation rate $u_j << 1$.

The waiting time for each step j is exponential, $exp(u_j)$.

⇒ The waiting time until stage k is reached:

$$T_k \sim exp(u_1) + \cdots + exp(u_k)$$

$$\mathbb{E}[T_k] = \mathbb{E}[\sum_{j=1}^{k} exp(u_j)] = \sum_{j=1}^{k} \frac{1}{u_j}$$

(2) Independent mutations :

Assume that each mutation occurs independently at time

$$T_j \sim exp(\lambda_j) \ , \ j=1, \dots, d$$

⇒ The waiting time until any k out of d mutations have occurred :

$$T_k \sim \underbrace{\min_{\{j_1, \dots, j_k\} \subset \{1, \dots, d\}}}_{\text{pick the earliest combination of } k} \underbrace{\max\{T_{j_1}, \dots, T_{j_k}\}}_{\text{waiting time for all } k \text{ mutations}}$$

$$T_1 \sim \min\{T_1, \dots, T_d\} = exp(\sum_{j=1}^{d} \lambda_j)$$

(3) Independent mutations + identical mutation rates :

Assume that

$$T_j \sim exp(\lambda) \ , \ j=1, \dots, d$$

Then, the waiting time until any k out of d mutations :

$$T_1 \sim exp(d\lambda)$$

$$T_k \sim \underset{\substack{\downarrow \\ \text{waiting time until} \\ \text{any } (k-1) \text{ out of } d \text{ mutations}}}{T_{k-1}} + \underbrace{exp((d-(k-1))\lambda)}_{\text{waiting time until any one out of } d-(k-1) \text{ mutations}}$$

$$\begin{aligned} \mathbb{E}[T_k] &= \mathbb{E}[T_{k-1}] + \frac{1}{(d-(k-1))\lambda} \\ &= \mathbb{E}[T_{k-2}] + \frac{1}{(d-(k-1))\lambda} + \frac{1}{(d-(k-2))\lambda} \\ &\ \vdots \\ &= \frac{1}{\lambda} \sum_{j=1}^{k} \frac{1}{d-j+1} \end{aligned}$$

(4) Mutational Pathways

- A mutational pathway is a genotype lattice can be written as

$$P = j_1 \to j_2 \to \cdots \to j_K$$

where $j_1 < \cdots < j_K$ defines the order of mutations.

- $Exit_i$:= set of all possible mutations in step i.
- The probability of $P = j_1 \to \cdots \to j_K$ is

$$\text{Prob}(P) = \prod_{i=1}^{K} \frac{\lambda_{j_i}}{\sum_{j \in Exit_i} \lambda_j}$$

- The expected waiting time is

$$\mathbb{E}[\tau_P] = \sum_{i=1}^{K} \frac{1}{\sum_{j \in Exit_i} \lambda_j}$$

- The expected waiting time of any K out of d mutations :

$$\mathbb{E}[\tau_K] = \underbrace{\sum_{P = j_1 \to \cdots \to j_K}}_{\text{sum over all possible pathways of length } K} \text{Prob}(P) \cdot \mathbb{E}[\tau_P]$$

② The Wright-Fisher Process

- Process : Consider a constant population size N and discrete generations. In each generation t, each individual chooses randomly a parent from the previous generation. The individuals can be of different types.

- Model :

Consider 2 types of individuals A and B.

Let $X(t)$:= # of type A individual in generation $t = 0, 1, 2, \ldots$

$$X(t) \in \{0, 1, \ldots, N\}$$

$$\Rightarrow X(t+1) \mid X(t) = i \sim \text{Binomial}(\underset{\substack{\downarrow \\ \#\text{ of draws}}}{N}, \underset{\substack{\downarrow \\ \text{proportion of type } A \text{ individuals}}}{\tfrac{i}{N}})$$

$$\Rightarrow P_{ij} = P(X(t+1) = j \mid X(t) = i)$$
$$= \binom{N}{j} \left(\frac{i}{N}\right)^j \left(1 - \frac{i}{N}\right)^{N-j}$$

- Stationary mean :

$$\mathbb{E}[X(t) \mid X(0) = i] = i$$

* Both the Moran process and the Wright-Fisher process model the genetic random drift.

Time-dependent Variance :

$$\lim_{N\to\infty} \mathrm{Var}(X(t) \mid X(0) = i) = i\left(1 - \frac{i}{N}\right) t$$

↳ ratio between WF process and Moran process $= \frac{N}{2}$

↳ Variance increases faster for the WF process because it updates all N individuals simultaneously, whereas the Moran process updates one at a time

- Heterozygosity :

$$\mathbb{E}[H_t \mid X(0) = i] = H_0(i)\left(1 - \frac{1}{N}\right)^t, \quad H_0(i) = 2 \cdot \frac{i}{N} \cdot \frac{N-i}{N-1}$$

↳ the heterozygosity decays exponentially at rate $\frac{1}{N} > \frac{2}{N^2}$, which is faster than in the Moran process.

- Fixation Probabilities :

2 absorbing states : $X(t) = 0$ and $X(t) = N$

Let the probability of fixation of type A individuals when starting with i be

$$x_i = \lim_{t\to\infty} \mathbb{P}(X(t) = N \mid X(0) = i)$$

We have

$$i = \lim_{t\to\infty} \mathbb{E}[X(t) \mid X(0) = i] = \overset{\text{only 2 possibilities}}{0 \times (1 - x_i) + N \times x_i}$$

$$\Leftrightarrow \qquad x_i = \frac{i}{N}$$

3) Wright-Fisher Process with Mutation and Selection

(1) Accumulating Mutations

Consider a binary genome of length d.

↳ each locus undergoes independent mutation from 0 to 1 at rate u

↳ no back mutations

Let $X_j(t) :=$ # of j-cells (cells with j mutations) in generation t

↳ $x_j(t) = \frac{X_j(t)}{N}$

Assume a constant fitness advantage s per mutation. → dominates the waiting time to cance

↳ the fitness of a j-cell $= (1+s)^j$

Assume $X_0(0) = N$, $X_j(0) = 0 \quad \forall j = 1, \dots, d$.

⇒ probability of sampling a j-cell

$$\theta_j(t) = \sum_{i=0}^{j} \mathbb{P}(i\text{-cell} \to j\text{-cell})$$

$$= \sum_{i=0}^{j} P(i\text{-to-}j \text{ mutations})\, \mathbb{P}(i\text{-cell parent})$$

$$= \sum_{i=0}^{j} \binom{d-i}{j-i} u^{j-i} (1-u)^{d-j} \cdot \frac{(1+s)^i x_i(t)}{\sum_\ell (1+s)^\ell x_\ell(t)}$$

The sampling step follows a multinomial distribution with parameters

- $X(t) = (X_0(t), \dots, X_d(t))$
- $\theta(t) = (\theta_0(t), \dots, \theta_d(t))$

Chapter 6 Diffusion Theory

① Structure of the Diffusion Theory

- Goal : model the probability distribution of a population over time
- Consider only 2 populations A_1 and A_2.
- $\psi(p,t)$:= probability density that A_1 has frequency p at time t
- $g(p,\varepsilon;dt)$:= probability that allele frequency of A_1 changes from p to $p+\varepsilon$ in time interval t

$$\Rightarrow \psi(p,t+dt) = \underbrace{\int \psi(p-\varepsilon,t)\,g(p-\varepsilon,\varepsilon;dt)\,d\varepsilon}_{\text{marginalizing out the amount of change } \varepsilon}$$

Interpretation : if the allele frequency is p at time $t+dt$, then it must have been $p-\varepsilon$ at time t for some amount $\varepsilon > 0$

② Two classes of Evolutionary Forces

(1) Directional processes : $M(p)$

↳ expected change in allele frequency per generation

↳ e.g. mutation, selection, migration, recombination

(2) Nondirectional processes : $V(p)$

↳ expected variance in the next generation

↳ e.g. genetic drift.

- example : Moran process

$$p = p(t) = \frac{X(t)}{N}$$

$$\Rightarrow M(p) = \mathbb{E}[p(t+1) - p(t) \mid p(t)] = p - p = 0$$

$$V(p) = \mathbb{E}[\,Var(p(t+1)) \mid p(t)\,] = \frac{1}{N^2}\mathbb{E}[Var(X(t+1)) \mid X(t)] = \frac{2p(1-p)}{N^2}$$

↳ the neutral Moran process models only the random drift

③ Kolmogorov Forward Equation / Fokker-Planck Equation / Diffusion Equation

$$\frac{\partial \psi(p,t)}{\partial t} = -\frac{\partial}{\partial p}[\psi(p,t)\, M(p)] + \frac{1}{2}\frac{\partial^2}{\partial p^2}[\psi(p,t)\, V(p)] \qquad (*)$$

- Derivation : Let $\psi := \psi(p,t)$, $g := g(p, \varepsilon; dt)$

$$\psi(p,t+dt) = \int \psi(p-\varepsilon, t)\, g(p-\varepsilon, \varepsilon; dt)\, d\varepsilon$$

$$= \int \left[\psi g - \varepsilon \frac{\partial(\psi g)}{\partial p} + \frac{\varepsilon^2}{2}\frac{\partial^2(\psi g)}{\partial p^2} - \frac{\varepsilon^3}{6}\frac{\partial^3(\psi g)}{\partial p^3} + \cdots\right] d\varepsilon \quad \text{by Taylor's expansion}$$

$$\approx \int \left[\psi g - \varepsilon \frac{\partial(\psi g)}{\partial p} + \frac{\varepsilon^2}{2}\cdot\frac{\partial^2(\psi g)}{\partial p^2}\right] d\varepsilon \quad \text{by assuming } \varepsilon^2 >> \varepsilon^3$$

$$= \psi \underbrace{\int g\, d\varepsilon}_{=1} - \frac{\partial}{\partial p}\psi \underbrace{\int g \varepsilon\, d\varepsilon}_{= M(p)\, dt} + \frac{1}{2}\frac{\partial^2}{\partial p^2}\psi \underbrace{\int g \varepsilon^2\, d\varepsilon}_{= V(p)\, dt} \quad \text{since } p \perp\!\!\!\perp \varepsilon$$

$$= \psi(p,t) - \frac{\partial[\psi(p,t)\, M(p)]}{\partial p}\, dt + \frac{1}{2}\frac{\partial^2[\psi(p,t)\, V(p)]}{\partial p^2}\, dt$$

↳ subtract both sides by $\psi(p,t)$

↳ divide by dt and let $dt \to 0$

- Equilibrium :

Setting (*) to zero gives

$$\frac{1}{2}\frac{\partial}{\partial p}[\psi^*(p,t)\, V(p)] - \psi^*(p,t)\, M(p) = 0$$

$$\Rightarrow \qquad \psi^*(p) = \frac{C}{V(p)} \exp\left(\int_0^p \frac{2M(q)}{V(q)}\, dq\right)$$

↳ In one-dimensional case :

$$\psi^*(p,t) = \frac{1}{\sqrt{2\pi v t}} \exp\left(-\frac{(p-mt)^2}{2vt}\right)$$

which is the pdf of $\mathcal{N}(mt, vt)$. As $t \to 0$, $\psi(p,t)$ converges to a point mass $p=0$

④ Combining Mutation, Selection, and Wright-Fisher type Sampling

(1) Selection

Assume A_1 and A_2 have frequencies p and $1-p$, with fitness w_1 and w_2

Then, the average fitness of the population is

$$\bar{w} = pw_1 + (1-p)w_2 \quad \text{and} \quad \frac{d\bar{w}}{dp} = w_1 - w_2$$

In the Wright-Fisher process, we sample a parent at random.

⇒ the allele frequency of A_1 in the next generation is

$$p' = \frac{pw_1}{pw_1 + (1-p)w_2} = \frac{pw_1}{\bar{w}}$$

⇒ difference in allele frequency due to selection :

$$\begin{aligned}\Delta p_{sel} &= p' - p \\ &= \frac{p(w_1 - \bar{w})}{\bar{w}} \quad \text{by definition} \\ &= \frac{p(1-p)(w_1 - w_2)}{\bar{w}} \quad \text{by definition} \\ &= p(1-p) \cdot \frac{1}{\bar{w}} \cdot \frac{d\bar{w}}{dp} \\ &= p(1-p) \frac{d \log(\bar{w})}{dp}\end{aligned}$$

↳ also called : Wright's equation for an adaptive landscape

(2) Mutation

Let $u_1 :=$ A_1-to-A_2 mutation rate

$u_2 :=$ A_2-to-A_1 mutation rate

$$\Rightarrow \quad \Delta p_{mut} = -pu_1 + (1-p)u_2$$

Combining selection and mutation we get :

$$M(p) = \underbrace{p(1-p)\frac{d\log(\bar{w})}{dp}}_{\Delta p_{sel}} \underbrace{- pu_1 + (1-p)u_2}_{\Delta p_{mut}}$$

(3) Sampling :

In the Wright-Fisher process :

$$Var[X(t+1) \mid X(t) = i] = Np(1-p) \quad \text{with } p = \frac{i}{N}$$

Thus,

$$\begin{aligned}V(p) &= \mathbb{E}[Var(p(t+1)) \mid p(t)] \\ &= \mathbb{E}[Var(\tfrac{1}{N} X(t+1)) \mid \tfrac{1}{N} X(t)] \\ &= \frac{1}{N^2} \cdot Np(1-p) \\ &= \frac{p(1-p)}{N}\end{aligned}$$

$$\begin{aligned}\Rightarrow \int_0^p \frac{M(q)}{V(q)} &= \int_0^p N\left(\frac{d\log(\bar{w})}{dq} - \frac{u_1}{1-q} + \frac{u_2}{q}\right) dq \\ &= N(\log \bar{w} + u_1 \log(1-p) + u_2 \log(p))\end{aligned}$$

$$\begin{aligned}\Rightarrow \psi^*(p) &= \frac{C}{V(p)} \exp\left(\int_0^p \frac{2M(q)}{V(q)} dq\right) \\ &= \frac{CN}{p(1-p)} \exp(2N(\log \bar{w} + u_1 \log(1-p) + u_2 \log(p))) \\ &= C \bar{w}^{2N} (1-p)^{2Nu_1 - 1} p^{2Nu_2 - 1}\end{aligned}$$

where C is a normalizing constant s.t. $\int_0^1 \psi^*(p) dp = 1$

(4) Equilibrium under mutation and drift

For $\bar{w} = 1$ and $u = u_1 = u_2$,

$$\psi^*(p) \propto [p(1-p)]^{2Nu - 1}$$

where $\theta = 2Nu$: scaled mutation parameter, which captures the impacts of both mutation rate and the population size on the distribution of allele frequency at equilibrium.

↳ for high mutation rate relative to N: coexistence ("unimodal")

↳ for low mutation rate relative to N: fixation to one of the two absorbing states

(5) Equilibrium under selection and drift

For $w_1 = 1 + s$ and $w_2 = 1$ for some small s and $u_1 = u_2 = 0$,

$$\Rightarrow \quad \bar{w} = p(1+s) + (1-p)$$
$$= 1 + sp$$
$$\approx e^{sp} \quad \text{for small } s$$

$$\Rightarrow \quad \psi^*(p) \propto \frac{\bar{w}^{2N}}{p(1-p)} \approx \frac{e^{2Nsp}}{p(1-p)}$$

where $\sigma = 2Ns$: scaled selection parameter, which captures the impacts of both selective advantage and population size on $\psi^*(p)$

↳ for large selective advantage : $\psi^*(p)$ concentrates on large p

↳ for small selective advantage : $\psi^*(p)$ concentrates on $\frac{1}{2}$

Putting (4) and (5) together gives the diffusion approximation for the Wright-Fisher process

$$\psi^*(p) \propto [p(1-p)]^{\theta - 1} e^{\sigma p}$$

with $\theta = 2Nu$, $\sigma = 2Ns$

(6) Compare with the Moran Process with constant selection

$$M_{WF}(p) \approx N \cdot M_{Moran}(p)$$

$$V_{WF}(p) \approx \frac{N}{2} \cdot V_{Moran}(p)$$

$$P_{WF}\left(\tfrac{1}{N}\right) = \frac{1 - e^{-2s}}{1 - e^{-2Ns}} \approx 2s \quad , \quad P_{Moran}\left(\tfrac{1}{N}\right) = \frac{1 - e^{-s}}{1 - e^{-Ns}} \approx s$$

⑤ Kolmogorov Backward Equation

$$\frac{\partial \psi(p,t|p_0)}{\partial t} = M(p_0)\frac{\partial \psi(p,t|p_0)}{\partial p_0} + \frac{1}{2}V(p_0)\frac{\partial^2 \psi(p,t|p_0)}{\partial p_0^2}$$

- Derivation :

Let $\psi := \psi(p,t|p_0) :=$ probability density that A_1 has allele frequency at time t, given that the allele frequency was p_0 at time 0.

$$\psi(p,t+dt|p_0) = \int \underbrace{\psi(p,t|p_0+\varepsilon)}_{\substack{\text{probability that } A_1 \\ \text{has frequency } p \text{ at time} \\ t+dt \text{ given that its} \\ \text{frequency is } p_0+\varepsilon \text{ at time } dt}} \underbrace{g(p_0,\varepsilon;dt)}_{\substack{\text{probability to go from} \\ p_0 \text{ to } p_0+\varepsilon \text{ in the interval } dt}} d\varepsilon$$

$$= \int \left[\psi(p,t|p_0) + \varepsilon\frac{\partial\psi}{\partial p_0} + \frac{\varepsilon^2}{2}\frac{\partial^2\psi}{\partial p_0^2} + \frac{\varepsilon^3}{6}\frac{\partial^3\psi}{\partial p_0^3} + \cdots\right] g\, d\varepsilon \quad \text{by Taylor's expansion}$$

$$\approx \int \left[\psi g + \varepsilon g\frac{\partial\psi}{\partial p_0} + \frac{1}{2}\varepsilon^2 g\frac{\partial^2\psi}{\partial p_0^2}\right] d\varepsilon \quad \text{assuming } \varepsilon^2 >> \varepsilon^3$$

$$= \psi \underbrace{\int g\, d\varepsilon}_{=1} + \frac{\partial\psi}{\partial p_0}\underbrace{\int \varepsilon g\, d\varepsilon}_{M(p_0)dt} + \frac{1}{2}\frac{\partial^2\psi}{\partial p_0^2}\underbrace{\int \varepsilon^2 g\, d\varepsilon}_{V(p_0)dt}$$

$$= \psi(p,t|p_0) + \left(M(p_0)\frac{\partial\psi(p,t|p_0)}{\partial p_0} + \frac{1}{2}V(p_0)\frac{\partial^2\psi(p,t|p_0)}{\partial p_0^2}\right)dt$$

- Equilibrium :

$$\frac{\partial\psi^*}{\partial p_0} = C\exp\left(-\int_0^p \frac{2M(q)}{V(q)}dq\right)$$

⑥ Fixation probabilities :

$P(p_0) = \psi(1,\infty|p_0) :=$ the fixation probability of A_1 given its initial frequency p_0

$\Rightarrow$ 2 absorbing states :

$$P(0) = \psi(1,\infty|0) = 0, \quad P(1) = \psi(1,\infty|1) = 1$$

$$\Rightarrow \quad P(p_0) = \frac{\int_0^{p_0}\exp\left(-\int_0^p \frac{2M(q)}{V(q)}dq\right)dp}{\int_0^1 \exp\left(-\int_0^p \frac{2M(q)}{V(q)}dq\right)dp}$$

- Wright-Fisher Process with Constant Selection :

With $w_1 = 1+s$ and $w_2 = 1$,

$$\bar{w} = 1+sp, \quad \frac{d\bar{w}}{dp} = s$$

$$\Rightarrow \quad M(p) = p(1-p)\frac{1}{\bar{w}}\cdot\frac{d\bar{w}}{dp} = \frac{sp(1-p)}{1+sp}$$

$$\Rightarrow \quad \frac{2M(p)}{V(p)} = \frac{2sp(1-p)}{1+sp} \Big/ (Np(1-p)) = \frac{2Ns}{1+sp}$$

$$\approx 2Ns \quad \text{for small } p$$

$$\Rightarrow \quad \int_0^p \frac{2M(q)}{V(q)}dq = 2Nsp$$

$$\Rightarrow \quad P(p_0) = \frac{1-e^{-2Nsp_0}}{1-e^{-2Ns}}$$

$$\lim_{s\to 0} P(p_0) = p_0$$

$$P\left(\frac{1}{N}\right) = \frac{1-e^{-2s}}{1-e^{-2Ns}} \Rightarrow P\left(\frac{1}{N}\right) \approx 2s$$

for large N, small s

7) Mean Fixation Time

Let $\tau(p_0)$:= expected waiting time until the fixation of A_1, given that it will be fixed and its initial frequency p_0

For the neutral Wright-Fisher process :

$$\tau\left(\frac{1}{N}\right) \cong 2N$$

If the mutant has an selective advantage/disadvantage s, the larger s is, the smaller $\tau(\frac{1}{N})$ ⇒ shorter waiting time for selected mutant

Chapter 7 Evolutionary Game Theory

① Frequency-Dependent Selection (Assuming infinite population)

- Idea: a selective advantage may decrease with increasing population size
- Consider 2 types A and B with frequencies

$$x_A(t) \quad \text{and} \quad x_B(t) \quad , \quad x(t) = (x_A(t), x_B(t))^T$$

and fitness

$$f_A(x(t)) \quad \text{and} \quad f_B(x(t))$$

- The average fitness is

$$\phi(x) = x_A f_A(x) + x_B f_B(x)$$

and the system is

$$\begin{cases} \dot{x}_A = x_A(f_A(x) - \phi(x)) \\ \dot{x}_B = x_B(f_B(x) - \phi(x)) \end{cases}$$

With $x := x_A$ and $1-x := x_B$, the system is equivalent as

$$\dot{x} = x(1-x)(f_A(x) - f_B(x))$$

- Equilibrium points:

(1) $x^* = 0$: stable if $f_A(0) < f_B(0)$

(2) $x^* = 1$: stable if $f_A(1) > f_B(1)$

(3) $x^* \mid x^* \in (0,1),\ f_A(x^*) = f_B(x^*)$: stable if

$$\frac{\partial f_A(x^*)}{\partial x} < \frac{\partial f_B(x^*)}{\partial x} \iff \underbrace{\frac{\partial}{\partial x}(f_A(x^*) - f_B(x^*)) < 0}_{\text{negative slope}}$$

② Evolutionary Game Theory

- Definition: study of frequency-dependent selection
- Setup:

(1) A population of players / individuals

(2) Fixed strategy

(3) Random interaction

(4) Goal is to increase the reproductive success / fitness

- Two-Player Game

$$\begin{array}{c|cc} & A & B \\ \hline A & a & b \\ B & c & d \end{array}$$

← A gets payoff

← B gets payoff

↑ Playing against A ↑ Playing against B

$$\Rightarrow \dot{x} = x(1-x)[\underbrace{(a-b-c+d)x+(b-d)}]$$

$$= \underbrace{ax + b(1-x)}_{f_A(x)} - \underbrace{cx - d(1-x)}_{-f_B(x)}$$

- Strategies :

(1) $a > c,\ b > d$: A is always the best strategy (A Nash)

(2) $a < c,\ b < d$: B is always the best strategy (B Nash)

(3) $a > c,\ b < d$: playing the same strategy as the opponent is always the best

$$\underset{\substack{\downarrow \\ \text{unstable}}}{x^*} = \frac{d-b}{a-c-b+d} \quad \text{and} \quad \frac{\partial}{\partial x}\left(f_A(x^*) - f_B(x^*)\right) = (a-c)+(d-b) > 0$$

(Both Nosh)

(4) $a < c,\ b > d$: playing the opposite stategy as the opponent is always the best

$$\underset{\substack{\downarrow \\ \text{stable}}}{x^*} = \frac{d-b}{a-c-b+d} \quad \text{and} \quad \frac{\partial}{\partial x}\left(f_A(x^*) - f_B(x^*)\right) = (a-c)+(d-b) < 0$$

(Both not Nash)

(5) $a = c,\ b = d$: same payoffs for all strategies

③ Nash Equilibrium :

If two playes play the same strategy and neither player can increase its payoff by changing strategy, then the strategy is at Nash equilibrium.

④ Evolutionary stable strategy (ESS)

We say that the selection of an all-A population will oppose the invasion of an infinitesimally small amount ε of B if

$$f_A(1-\varepsilon) > f_B(\varepsilon)$$

$$\Leftrightarrow \quad a(1-\varepsilon) + b\varepsilon > c(1-\varepsilon) + d\varepsilon$$

A is ESS if either $\underset{\substack{\downarrow \\ \text{A strictly Nash}}}{a > c}$ or $\underset{\substack{\downarrow \\ \text{A Nash}}}{a = c}$ and $\underset{\substack{\downarrow \\ \text{B not Nash}}}{b > d}$

⑤ n-Player Evolutionary Game

Suppose there are n strategies.

$$\begin{array}{c|cccc} & S_1 & S_2 & \cdots & S_n \\ \hline S_1 & a_{11} & a_{12} & \cdots & a_{1n} \\ S_2 & a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ S_n & a_{n1} & a_{n2} & \cdots & a_{nn} \end{array}$$

The payoff of playing strategy S_i against strategy S_j is

$$E(S_i, S_j) = a_{ij}$$

(1) Definitions for strategies

S_K is unbeatable if $\forall\, i \neq k$, $a_{kk} > a_{ik}$ and $a_{ki} > a_{ii}$

$\Downarrow$

S_K is strict Nash if $\forall\, i \neq k$, $a_{kk} > a_{ik}$

$\Downarrow$

S_K is ESS if $\forall\, i \neq k$, either $a_{kk} > a_{ik}$ or $a_{kk} = a_{ik}$ and $a_{ki} > a_{ii}$

$\Downarrow$

S_K is weak ESS if $\forall\, i \neq k$, either $a_{kk} > a_{ik}$ or $a_{kk} = a_{ik}$ and $a_{ki} \geq a_{ii}$

$\Downarrow$

S_K is Nash if $\forall\, i \neq k$, $a_{kk} \geq a_{ik}$

* Only the first four are stable against invasion (Evolutionary Stable)

(2) Fitness = Expected Payoff

$$f_i(x) = f_{S_i}(x) = \sum_{j=1}^{n} x_j a_{ij}$$

$$\Rightarrow \quad \phi(x) = \sum_{i=1}^{n} x_i f_i(x)$$

(3) Replicator Equation

$$\dot{x}_i = x_i \left(f_i(x) - \phi(x) \right), \quad i = 1, \dots, n$$

where $x = (x_1, \dots, x_n)^T$, $\underbrace{x_1 + \cdots + x_n = 1}_{\text{simplex } S_n}$

↳ the interior of S_n and each face are invariant

↳ vertices of S_n are fixed points

(4) 3-Player Game / RPC Game

$$A = \begin{pmatrix} 0 & -a_2 & b_3 \\ b_1 & 0 & -a_3 \\ -a_1 & b_2 & 0 \end{pmatrix}$$

- $\det(A) > 0$: unique interior equilibrium, globally stable (damped oscillations)
- $\det(A) < 0$: unique interior equilibrium, unstable (increasing oscillations)
- Special case : zero-sum game

$$A = \begin{pmatrix} 0 & -1 & 1 \\ 1 & 0 & -1 \\ -1 & 1 & 0 \end{pmatrix}$$

↳ average fitness is always zero $\Rightarrow$ $\dot{x}_i = x_i f_i(x)$

(5) $n \geq 4$: allows for limit cycles and chaotic attracters

↳ at most one isolated equilibrium in the interior : $f_1 = \cdots = f_n$, $x_1 + \cdots + x_n = 1$

6) Hawks and Doves

Consider 2 stategies : Hawks (H) escalate fights / Doves (D) retreat.

Benefit of winning = b vs. cost of injury = c

$$\begin{array}{c|cc} & H & D \\ \hline H & \frac{b-c}{2} & b \\ D & 0 & \frac{b}{2} \end{array}$$

⇒ Replicator equation :

$$\dot{x}_H = x_H(1-x_H)\left[\frac{b-c}{2}x_H + b(1-x_H) - \frac{b}{2}(1-x_H)\right]$$

$$= x_H(1-x_H)\left(-\frac{c}{2}x_H + \frac{b}{2}\right)$$

⇒ Interior equilibrium :

$$x_H^* = \frac{b}{c}$$

↳ which is stable if $b < c$ ⇒ hawks and doves can coexist

- Mixed Strategies

Consider a strategy that plays H with prob. p and D with prob. $(1-p)$.

The payoff of playing strategy p_1 against strategy p_2 is

$$E[p_1, p_2] = p_1 p_2 E[H,H] + p_1(1-p_2)E[H,D] + (1-p_1)p_2 E[D,H] + (1-p_1)(1-p_2)E[D,D]$$

$$= p_1p_2\frac{b-c}{2} + p_1(1-p_2)b + (1-p_1)(1-p_2)\frac{b}{2}$$

$$= \frac{b}{2}p_1p_2 - \frac{b}{2}\cdot\frac{c}{b}p_1p_2 + \frac{b}{2}\cdot 2p_1 - \frac{b}{2}\cdot 2p_1p_2 + \frac{b}{2} - \frac{b}{2}p_1 - \frac{b}{2}p_2 + \frac{b}{2}p_1p_2$$

$$= \frac{b}{2}\left(1 + p_1 - p_2 - \frac{c}{b}p_1p_2\right)$$

The strategy $p^* = \frac{b}{c}$ is evolutionary stable since

$$\left.\begin{aligned} E[p^*, p^*] &= \frac{b}{2}\left(1 - \frac{b}{c}\right) \\ E[p^*, p] &= \frac{b}{2}\left(1 + \frac{b}{c} - 2p\right) \\ E[p, p^*] &= \frac{b}{2}\left(1 - \frac{b}{c}\right) \\ E[p, p] &= \frac{b}{2}\left(1 - \frac{c}{b}p^2\right) \end{aligned}\right\} \Rightarrow \begin{aligned} &E[p^*, p^*] = E[p, p^*] \\ &E[p^*, p] > E[p, p] \;\; \forall p \neq p^* . \end{aligned}$$

⇒ p^* is Nash, weak ESS, ESS but not strictly Nash or unbeatable

- If we consider only pure strategies, then there may not be a Nash equilibrium. But if we consider all mixed strategies, there is always a Nash equilibrium.

1) Prisoner's Dilemma

- 2 strategies : (1) Cooperation (C) , (2) Defection (D)

$$\begin{array}{c|cc} & C & D \\ \hline C & R & S \\ D & T & P \end{array}$$

$T > R > P > S$ and $R > \frac{T+P}{2}$

- CC : Reward for mutual cooperation
- DC : Temptation to defect
- DD : Punishment for mutual defection
- CD : Sucker's payoff

- Direct Reciprocity : the game is repeated m times

(1) GRIM vs ALLD :

- GRIM : cooperate until the opponent defects
- ALLD : always defect

$$\begin{array}{c|cc} & \text{GRIM} & \text{ALLD} \\ \hline \text{GRIM} & mR & S+(m-1)P \\ \text{ALLD} & T+(m-1)P & mP \end{array}$$

For $m > \frac{T-P}{R-P}$, GRIM and ALLD are both strictly Nash

$\Rightarrow$ Direct Reciprocity can stabilize cooperation but not initiate.

(2) GRIM vs GRIM* :

- GRIM* : given m , defect on the mth round

$$\begin{array}{c|cc} & \text{GRIM} & \text{GRIM}^* \\ \hline \text{GRIM} & mR & (m-1)R+S \\ \text{GRIM}^* & (m-1)R+T & (m-1)R+P \end{array}$$

$\Rightarrow$ GRIM $\rightarrow$ GRIM* $\rightarrow$ GRIM** $\rightarrow$ ... $\rightarrow$ ALLD : only Nash equilibrium

- Variable number of rounds

(1) w := probability that another round will be played

$\Rightarrow$ probability of playing k rounds : $w^{k-1}(1-w)$

$\Rightarrow$ expected # of rounds :

$$\bar{m} = \sum_{k=1}^{\infty} w^{k-1}(1-w) = \frac{1}{1-w}$$

$$\begin{array}{c|cc} & \text{GRIM} & \text{ALLD} \\ \hline \text{GRIM} & \bar{m}R & S+(\bar{m}-1)P \\ \text{ALLD} & T+(\bar{m}-1)P & \bar{m}P \end{array}$$

$\Rightarrow$ For $\bar{m} > \frac{T-P}{R-P}$, GRIM is evolutionary stable

$\Rightarrow$ defecting in the last round is no longer possible

(2) TFT vs. ALLD

- TFT : start with cooperation and do what the opponent have done in the previous round

$$\begin{array}{c|cc} & \text{TFT} & \text{ALLD} \\ \hline \text{TFT} & \bar{m}R & S+(\bar{m}-1)P \\ \text{ALLD} & T+(\bar{m}-1)P & \bar{m}P \end{array}$$

↳ same payoff matrix as GRIM vs. ALLD

↳ For $\bar{m} > \frac{T-P}{R-P}$, TFT is evolutionary stable against ALLD

↳ it can resume cooperation while GRIM cannot

↳ not robust against error, if one player starts to defect, then both defeet $\to \infty$

↳ TFT can be easily invaded by ALLC

(3) TFT vs. ALLC

- ALLC : always cooperate

$$\begin{array}{c|cc} & \text{TFT} & \text{ALLC} \\ \hline \text{TFT} & \bar{m}R & \bar{m}R \\ \text{ALLC} & \bar{m}R & \bar{m}R \end{array}$$

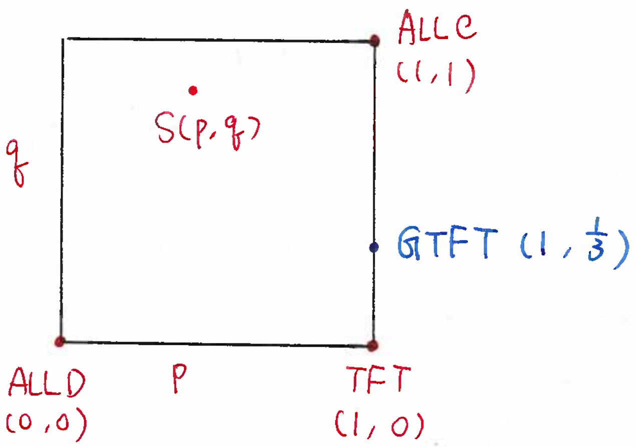
↳ TFT is not evolutionary stable against random drift.

3) Reactive Strategies

Strategy $S(p, q)$ cooperates with

(1) prob. p if the opponent cooperated in the previous move

(2) prob. q if the opponent defected in the previous move



- Markov chain for repeated Prisoner's Dilemma

Consider 2 reactive strategies $S_1(p_1, q_1)$ and $S_2(p_2, q_2)$

State space : $\{CC, CD, DC, DD\}$

Transition matrix : M

$$\begin{array}{c|cccc} & CC & CD & DC & DD \\ \hline CC & p_1 p_2 & p_1(1-p_2) & (1-p_1)p_2 & (1-p_1)(1-p_2) \\ CD & q_1 p_2 & q_1(1-p_2) & (1-q_1)p_2 & (1-q_1)(1-p_2) \\ DC & p_1 q_2 & p_1(1-q_2) & (1-p_1)q_2 & (1-p_1)(1-q_2) \\ DD & q_1 q_2 & q_1(1-q_2) & (1-q_1)q_2 & (1-q_1)(1-q_2) \end{array}$$

Probability distribution of the game after t rounds :

$$x(t) = (x_{CC}(t), x_{CD}(t), x_{DC}(t), x_{DD}(t))$$

Dynamical system :

$$x(t+1) = x(t) \cdot M$$

Payoff at stationary distribution :

$$E(S_1, S_2) = R s_1 s_2 + S s_1(1-s_2) + T(1-s_1)s_2 + P(1-s_1)(1-s_2)$$

where $s_1(p_1, p_2, q_1, q_2)$ and $s_2(p_1, p_2, q_1, q_2)$ are stationary distributions of cooperat

- Generous TFT (GTFT)

$$GTFT = S(1, q), \quad q = \min\left\{1 - \frac{T-R}{R-S}, \frac{R-P}{T-P}\right\}$$

$$ALLD \rightarrow TFT \rightarrow GTFT$$

because GTFT can correct mistakes stochastically

- Win-Stay Lose-Shift (WSLS)

Cooperate when CC or DD , defect when CD or DC

↳ deterministic corrector for errors

↳ WSLS dominates ALLC , whereas GTFT does not .

hapter 8 Evolutionary Games in Finite Populations

1) Two Players in a finite population

Consider 2 strategies : A , B

Population size : N

Payoff matrix :

$$\begin{array}{c|cc} & A & B \\ \hline A & a & b \\ B & c & d \end{array}$$

Let $i :=$ # of A individuals

The expected payoff for A and B :

freq of type B

$$A: \qquad F_i = \frac{(i-1)a + (N-i)b}{N-1}$$

freq of other type A

$$B: \qquad G_i = \frac{ic + (N-i-1)d}{N-1}$$

We say that selection opposes A invading B if

$$F_1 < G_1$$

$$\Leftrightarrow \quad (N-1)b < c + (N-2)d \text{ , independent of } a$$

For $N=2$, we have simply $b < c$.

↳ playing A against B gets smaller payoff than playing B against A

2) Intensity of Selection

Let $w :=$ intensity of selection.

Define the frequency-dependent fitness as

$$f_i = 1 - w + wF_i$$

$$g_i = 1 - w + wG_i$$

↳ if $w=0$, no force of selection / the game does not contribute

↳ if $w=1$, fitness is determined entirely by the payoff

↳ weak selection : send $w \to 0$

↳ w gets cancelled out in the deterministic replicator equation but is important in the stochastic process in finite population

3) Moran Process with Constant Selection

Replace the original fitness values by frequency-dependent fitness :

$$P_{i,i+1} = \underbrace{\frac{if_i}{if_i + (N-i)g_i}}_{P(T_A < T_B)} \cdot \underbrace{\frac{N-i}{N}}_{\text{choose a B to die}}$$

$$P_{i,i-1} = \underbrace{\frac{(N-i)g_i}{if_i + (N-i)g_i}}_{P(T_A > T_B)} \cdot \underbrace{\frac{i}{N}}_{\text{choose an A to die}}$$

$$P_{i,i} = 1 - P_{i,i+1} - P_{i,i-1}$$

and $P_{0,0} = P_{N,N} = 1$.

- Fixation Probability

$$\gamma_i = \frac{P_{i,i-1}}{P_{i,i+1}} \Rightarrow \rho_A = \frac{1}{1 + \sum_{j=1}^{N-1} \prod_{k=1}^{j} \frac{g_i}{f_i}} = \frac{1}{1 + \sum_{j=1}^{N-1} \prod_{k=1}^{j} \frac{1-w+wG_i}{1-w+wF_i}}$$

4) Weak Selection Limit

As $w \to 0$, $\rho_A \approx \frac{1}{N} \frac{1}{1-(\alpha N - \beta)w/6}$

where $\alpha = a+2b-c-2d$, $\beta = 2a+b+c-4d$

↳ selection favors the fixation of A if

$$\rho_A > \frac{1}{N} \iff \alpha N > \beta \iff a(N-2) + b(2N-1) > c(N+1) + d(2N-4)$$

↳ For $N=2$, we have

$$\rho_A > \frac{1}{N} \iff b > c \iff F_1 > G_1$$

- Weak Selection and Large Population

As $N \to \infty$, we have

$$\rho_A > \frac{1}{N} \iff a + 2b > c + 2d \iff a-c > 2(d-b)$$

In the case where $a > c$, $b < d$:

- Both A and B are Nash. Whichever has higher frequency gets higher fitness.
- The unstable interior equilibrium :

$$x^* = \frac{d-b}{a-b-c+d} = \frac{d-b}{(a-c)+(d-b)}$$

seems contradictory

$\Rightarrow$ The $\frac{1}{3}$ Law :

$$\rho_A > \frac{1}{N} \iff x^* < \frac{1}{3} \leftarrow$$

* it can happen that selection opposes the invasion of A but favours the fixation of A

5) Evolutionary Stability in Finite Population

$$\begin{array}{c|cc} & A & B \\ \hline A & a & b \\ B & c & d \end{array}$$

B is ESS_N if

(1) Selection protects against invasion:

$$F_1 < G_1 \iff (N-1)\,b < c + (N-2)\,d$$

(2) selection protects against replacement:

$$\rho_A < \frac{1}{N} \iff a(N-2) + b(2N-1) < c(N+1) + d(2N-4)$$

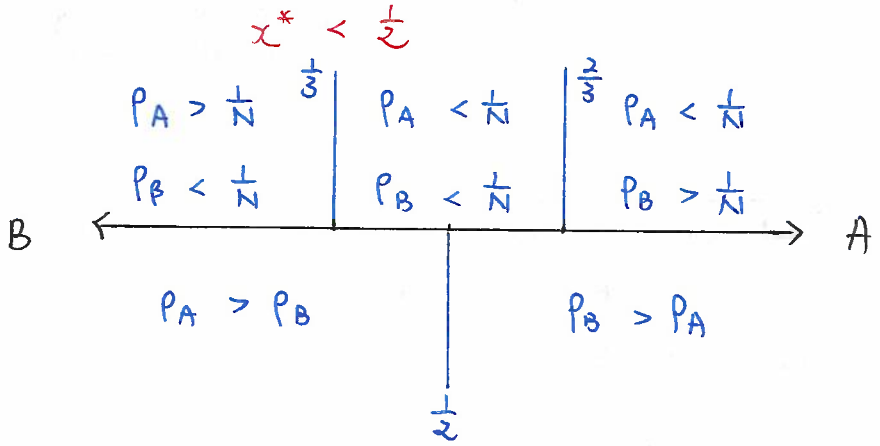
↳ For $N=2$, B is ESS_N if $b<c$, ESS is neither necessary nor sufficient

↳ For large N, B is ESS_N if $b<d$ and $x^* > \frac{1}{3}$, ESS is necessary but not sufficient.

6) Risk Dominance

A is risk dominant over B if $\rho_A > \rho_B$.

For large N, we have



7) Prisoner's Dilemma in Finite Population

- In the case of an infinite population, both TFT and ALLD are stable against invasion by each other if $m > \dfrac{T-P}{R-P}$
- In the finite population, selection favors TFT to replace ALLD if $\rho_{TFT} > \frac{1}{N}$

Chapter 8 Evolutionary Graph Theory

① Evolutionary Graph

- $G = (V, E)$ where $V = \{1, 2, \dots, N\}$ are individuals and $i \to j \in E$ means offspring of i can replace j
- In each generation, one individual $i \in \{1, 2, \dots, N\}$ is chosen randomly and its offspring replaces j with probability w_{ij}.

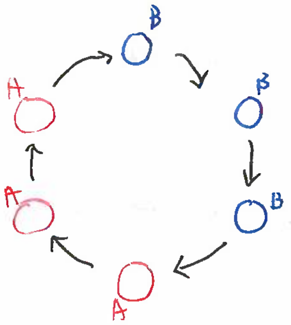
  The matrix $W = (w_{ij})$ is a stochastic matrix

- Moran process

$$w_{ij} = \frac{1}{N} \quad \forall\, i, j \quad , \quad \rho = \frac{1 - 1/r}{1 - 1/r^N}$$

↳ a complete graph with identical weights

- Markov chain on a directed cycle :



$$\begin{pmatrix} 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 & 0 & 0 \end{pmatrix}$$

↳ starting from one B mutant, only one connected cluster of B can emerge

↳ Let $m :=$ # of B individuals with fitness r. A has fitness 1.

$$\Rightarrow \quad P_{m, m+1} = \frac{r}{N - m + rm}$$

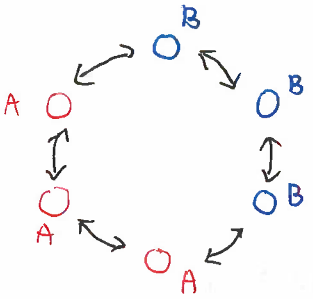
$$P_{m, m-1} = \frac{1}{N - m + rm}$$

$$\Rightarrow \quad \gamma_m = \frac{1}{r}$$

$$\Rightarrow \quad \rho = \frac{1 - 1/r}{1 - 1/r^N}$$ (same as the Moran process)

- Bidirected Cycle

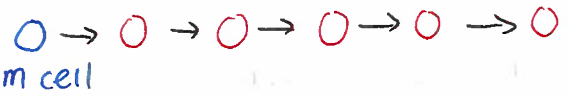
$$\begin{pmatrix} 0 & \frac{1}{2} & 0 & 0 & 0 & \frac{1}{2} \\ \frac{1}{2} & 0 & \frac{1}{2} & 0 & 0 & 0 \\ 0 & \frac{1}{2} & 0 & \frac{1}{2} & 0 & 0 \\ 0 & 0 & \frac{1}{2} & 0 & \frac{1}{2} & 0 \\ 0 & 0 & 0 & \frac{1}{2} & 0 & \frac{1}{2} \\ \frac{1}{2} & 0 & 0 & 0 & \frac{1}{2} & 0 \end{pmatrix}$$



$$\Rightarrow \rho = \frac{1 - 1/r}{1 - 1/r^N}$$

↳ Now, each cell has 2 directions to place offspring but only one direction will increase its frequency.

- Linear Process

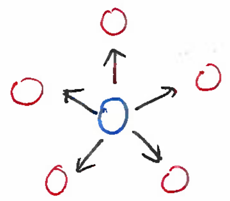


$$\begin{pmatrix} 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix}$$

↳ expected fixation of a randomly placed mutant :

$$\bar{\rho} = \frac{1}{N}$$

- Burst



$$\begin{pmatrix} 0 & \frac{1}{5} & \frac{1}{5} & \frac{1}{5} & \frac{1}{5} & \frac{1}{5} \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix}$$

↳ again, $\bar{\rho} = \frac{1}{N}$

② Amplifiers and Suppressors

- A graph G is an amplifier of selection if for $r > 1$

$$\rho_G > \rho_{Moran}$$

- A graph G is a suppressor of selection if for $r < 1$

$$\rho_G < \rho_{Moran}$$

- The strongest suppressor of selection is independent of r :

$$\rho_G = \frac{1}{N}$$

③ Isothermal Theorem

- The temperature of a vertex j :

$$T_j = w_{ij} + \cdots + w_{nj}$$

- Hot : many incoming edges ⇒ isothermal := all vertices have the same temperatu
- $\rho_G = \rho_{Moran} \iff G$ is isothermal $\iff G$ is doubly stochastic
- Isothermal graphs
  1) Directed cycle
  2) Cycle
  3) All symmetric graphs : $w_{ij} = w_{ji} \quad \forall i, j$
- All one-rooted graphs : $\rho = \frac{1}{N}$
- Multiple-rooted graphs prevent fixation
- Star (Bidirected burst) is an amplifier of selection
- Superstar is a strong amplifier of selection

④ Games on evolutionary graph (slides)

Chapter 9 Spatial Models of the Evolution of Solid Tumours

① Four processes of population genetics

(1) selection (2) mutation (3) drift (4) gene flow

↑

spatial structure

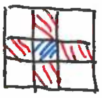
② Cellular Automata

- regular grid of sites, which are associated with sets of states
- each site is in a neighbourhood
- rules : depend on current state of site and its neighbourhood

  ↳ can be probabilistic

③ Eden Growth Model

- 2 states : unoccupied (S_0) , occupied (S_1)
- von Neumann neighbourhood : adjacent sites



- Available site-focused rule : (smooth boundary

  randomly choose an S_0 site that adjoins at least one S_1 site and switch to S_1
- Bond-focused rule :

  randomly choose an S_1 site with probability proportional to the number of S_0 sites then randomly choose an S_0 neighbour and switch to S_1 (rough boundary)
- Cell-focused rule :

  randomly choose an S_1 site that adjoins at least one S_0 site, then randomly choose an S_0 neighbour and switch to S_1 (smooth boundary)

↳ all resemble a disc in 2D and a ball in 3D in the long run

④ Eden Growth Model with Mutation

- Multiple occupied states $\{S_1, S_2, \ldots\}$
- Mutation rates : $S_i \to S_j$
- Mutation at division
- Neutral model : a disc
- With selection : e.g. $S_i : (1+s)^i$ ⇒ new mutations grow faster

) Eden Growth Model w̄ Cell Death and Migration

- Cell Death facilitates selection
- Migration facilitates growth by increasing the surface-to-volume ratio

) Deme-Based Models

- each deme contains multiple cells
- assume cells within demes are well-mixed
- cells can migrate between demes
- When each deme contains only one cell, back to cellular automaton
- As migration rate → 0, reduce to a set of independent non-spatial processes

7) Spatial Moran Model

- Assumptions:
  1) Offspring of one cell replaces another cell
  2) replacement is chosen with prob. ∝ cell fitness (death-birth model)
  3) prob. of being replaced by local parent = 1 − m
     " " " " " neighbouring " = m = dispersion probability

- Consider a mutant invading an infinite row of demes

  Let $n = \{\cdots, n_{i-1}, n_i, n_{i+1}, \cdots\}$ be the vector of mutant population sizes

  Let μ := death rate, s := fitness advantage, N := deme population size

  - Probability that the number of mutants n_i in deme i increases (decreases) by one:

$$W_i^+(n) = \underbrace{\mu(N-n_i)}_{\text{prob. that a WT cell dies}} \times \underbrace{(1+s)}_{\text{fitness of mutant}} \times \Big(\underbrace{(1-m)\times\frac{n_i}{N} + \frac{m}{2}\times\frac{n_{i-1}}{N} + \frac{m}{2}\times\frac{n_{i+1}}{N}}_{\text{replacement + dispersion}}\Big)$$

$$W_i^-(n) = \mu n_i \times \Big((1-m)\frac{N-n_i}{N} + \frac{m}{2}\times\frac{N-n_{i-1}}{N} + \frac{m}{2}\times\frac{N-n_{i+1}}{N}\Big)$$

- Diffusion Approximation / Fisher's Equation

$$\frac{\partial u}{\partial t} = \underbrace{D[1+s(1-u)]\frac{\partial^2 u}{\partial x^2}}_{\text{speed of mutant spreading through space}} + \underbrace{\mu s u(1-u)}_{\text{growth of a mutant within a deme}} \approx D\frac{\partial^2 u}{\partial x^2} + ru(1-u)$$

↳ $u = \frac{\langle n_i \rangle}{N}$, $x \approx l_i$ (distance along the row of demes)

↳ approximation only works when N is large and s is large

↳ should also consider clonal interference, environmental heterogeneity

Chapter 10 Branching Processes in Biology

① Galton-Watson Process

- Process: A single ancestor lives for one unit of time, after which it produces a random number of offspring $Z \sim P$ fixed. Each offspring i.i.d. $\sim P$

Let $Z_n :=$ # of individuals in generation n. $Z_0 = 1$ and $Z_1 = Z$.

The Galton-Watson process is

$$\{Z_n \mid n = 0, 1, 2, \dots\}$$

- defined on the nonnegetive integers

$$\Rightarrow \quad Z_{n+1} = \sum_{j=1}^{Z_1} Z_n^{(j)} \longrightarrow$$

# of offspring that ancestor j in generation 1 produces in generation n

$\downarrow$ random variable

- Transition Probabilities:

Let $p_k = \text{Prob}(Z = k)$, $k \in \mathbb{N}$

$$\Rightarrow \quad p(i,j) = \text{Prob}(Z_{n+1} = j \mid Z_n = i)$$

$$\Rightarrow \quad p(1,k) = p_k, \quad k \in \mathbb{N}$$

In general:

$$p(0,j) = \delta_{0j} = \begin{cases} 1 & \text{if } j = 0 \\ 0 & \text{else} \end{cases}$$

For $i \geq 1$:

$$p(i,j) = p_j^{*i} = \underbrace{\sum_{k_1 + k_2 + \dots + k_i = j}}_{\text{all combinations of producing } j \text{ offsprings}} p_{k_1} \cdots p_{k_i}$$

↳ $\{p_k^{*i}\}_{k \geq 0} :=$ i-fold convolution of $\{p_k\}_{k \geq 0}$

② Probability Generating Functions

For $Z \sim \{p_k\}_{k \geq 0}$, the pgf of Z is

$$f(s) = \mathbb{E}[s^Z] = \sum_{k=0}^{\infty} p_k s^k, \quad s \in [0,1] \quad \Rightarrow f(1) = \sum_{k=0}^{\infty} p_k = 1$$

The pgf generates the distribution p through derivatives:

$$\frac{d^k f}{ds^k}(0) = k!\, p_k, \quad k \geq 0$$

(1) Moments of Z:

$$f'(s) = \sum_{k=0}^{\infty} k p_k s^{k-1}, \quad f''(s) = \sum_{k=0}^{\infty} k(k-1) p_k s^{k-2} = \sum_{k=0}^{\infty} k^2 p_k s^{k-2} - \sum_{k=0}^{\infty} k p_k s^{k-2}$$

$$\Rightarrow \quad \mathbb{E}[Z] = f'(1), \quad Var(Z) = \mathbb{E}[Z^2] - (\mathbb{E}[Z])^2 = f'(1) + f''(1) - f'(1)^2$$

(2) Powers of f:

$$f(s) = \sum_{j=0}^{\infty} p_j s^j = \sum_{j=0}^{\infty} P(1,j)\, s^j$$

$$\Rightarrow [f(s)]^k = \sum_{j=0}^{\infty} P(k,j)\, s^j \quad , \quad k \geqslant 1$$

(3) Iterations:

$$f^{(0)}(s) = s$$

$$f^{(1)}(s) = f(s)$$

$$f^{(n+1)}(s) = f(f^{(n)}(s)) \quad , \quad n \geqslant 1$$

(4) n-step Transitions

$P_n(i,j) :=$ transition probability from i individuals to j individuals in n steps

(5) Chapman-Kolmogorov Equation

$$P_{n+m}(i,j) = \sum_{k=0}^{\infty} P_n(i,k)\, P_m(k,j)$$

(6) Proposition: $f_n = f^{(n)}$

Let f_n be the pgf of Z_n. Then f_n is equivalent as applying the pgf of z for n times.

proof: $f_{n+1}(s) = \sum_j P_{n+1}(1,j)\, s^j$ by definition of f_{n+1}

$= \sum_j \sum_k P_n(1,k) P(k,j)\, s^j$ by Chapman-Kolmogorov equation

$= \sum_k P_n(1,k) \sum_j P(k,j)\, s^j$ by rearranging

$= \sum_k P_n(1,k)\, [f(s)]^k$ by definition of power of f

$= f_n(f(s))$ by definition of f_n

$\vdots$

$= f^{(n+1)}(s)$ ▨

7) Moments of Z_n

Assume that $p_0 + p_1 < 1$ and $p_j \neq 1$ for all j

Set $m = \mathbb{E}[Z]$ and $\sigma^2 = Var[Z]$. Then

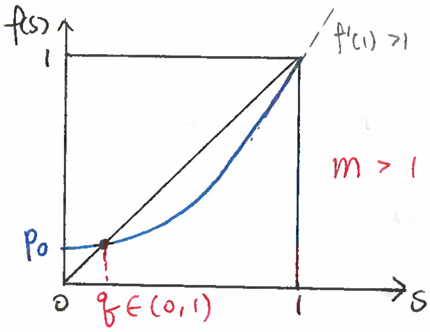
$$\mathbb{E}[Z_n] = \underset{\downarrow}{m^n} \quad , \quad Var[Z_n] = \begin{cases} \dfrac{\sigma^2 m^{n-1}(m^n - 1)}{m-1} & \text{if } m \neq 1 \\ n\sigma^2 & \text{if } m = 1 \end{cases}$$

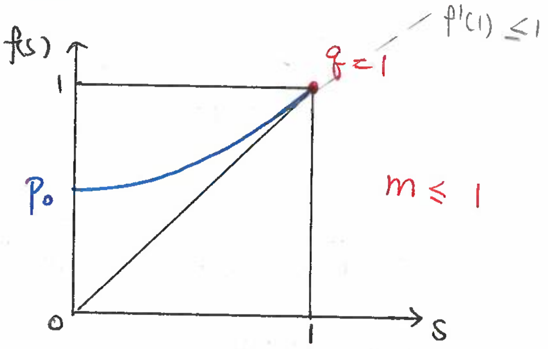
exponential in expected # of offsprings

(3) Extinction

Given that $Z_n = 0$ is an absorbing state, the probability of extinction is

$$p = \text{Prob}(Z_i = 0 \text{ for some } i \geq 0)$$
$$= \lim_{n\to\infty} \text{Prob}(Z_i = 0 \text{ for some } 1 \leq i \leq n)$$
$$= \lim_{n\to\infty} \text{Prob}(Z_n = 0)$$
$$= \lim_{n\to\infty} f_n(0) = \lim_{n\to\infty} f^{(n)}(0)$$





- Theorem

The extinction probability of the Galton-Watson process $\{Z_n\}$ is the smallest non-negative root q of the equation $f(s) = s$.

If $E[Z] = m \leq 1$, then $q = 1$. If $m > 1$, then $q < 1$.

- Criticality

(1) Supercritical

$m > 1$, $E[Z_n] \to \infty$, $q < 1$

(2) Critical

$m = 1$, $E[Z_n] = 1$, $q = 1$

(3) Subcritical

$m < 1$, $E[Z_n] \to 0$, $q = 1$

- Instability :

$$\lim_{n\to\infty} \text{Prob}(Z_n = k) = 0 \quad , \quad k \geq 1$$

$$\text{Prob}\left(\lim_{n\to\infty} Z_n = 0\right) = q$$

$$\text{Prob}\left(\lim_{n\to\infty} Z_n = \infty\right) = 1 - q$$

Interpretation : In the Galton-Watson process, the population cannot stay in any state indefinitely. It will either go extinct with probability q, or keep on growing forever with probability $(1-q)$ ⇒ useful for small population

④ Multi-Type Galton Watson Process

- Consider 2 types : type 0 (Wild type) , type 1 (mutant) with counts $Z_0(t)$ and $Z_1(t)$, $t \in \{0,1,2,\ldots\}$

- Each cell gives 2 offsprings.

- Each offspring of a type 0 cell can mutate to type 1 with rate α , irreversible

- Probability generating functions

$$F = (F_0, F_1)$$

$$F_0(s_0, s_1; t) = \mathbb{E}[s_0^{Z_0(t)} s_1^{Z_1(t)} \mid Z_0(0)=1 , Z_1(0)=0] := F_0(s;t)$$

$$F_1(s_0, s_1; t) = \mathbb{E}[s_0^{Z_0(t)} s_1^{Z_1(t)} \mid Z_0(0)=0 , Z_1(0)=1] := F_1(s;t)$$

$$\Rightarrow F_0(s;t) = [(1-\alpha)F_0(s;t-1) + \alpha F_1(s;t-1)]^2$$

$$F_1(s;t) = [F_1(s;t-1)]^2$$

$$\Rightarrow \mathbb{E}[Z_0(t) \mid Z_i(0) = \delta_{1i}] = 2\,\mathbb{E}[Z_0(t-1) \mid Z_i(0) = \delta_{1i}] = 0$$

$$\mathbb{E}[Z_0(t) \mid Z_i(0) = \delta_{0i}] = 2(1-\alpha)\,\mathbb{E}[Z_0(t-1) \mid Z_i(0) = \delta_{0i}]$$

$$\Rightarrow \mathbb{E}[Z_0(t) \mid Z_i(0) = \delta_{0i}] = [2(1-\alpha)]^t$$ = expected number of WT at time t

Expected total # of cells

$$N(t) = \mathbb{E}[Z_0(t) + Z_1(t) \mid Z_i(0) = \delta_{0i}] = 2^t$$

Expected # of mutant cells

$$r(t) = \mathbb{E}[Z_1(t) \mid Z_i(0) = \delta_{0i}] = 2^t - (2(1-\alpha))^t = 2^t(1-(1-\alpha)^t)$$

Probability of a mutant-free population

$$P_0(t) = F_0(1,0;t) = \mathbb{E}[1^{Z_0(t)} 0^{Z_1(t)} \mid Z_i(0) = \delta_{0i}] = \mathbb{E}[\mathbb{I}\{Z_1(t)=0\} \mid Z_i(0) = \delta_{0i}].$$

$P_1(t) = F_1(1,0;t)$ = prob. of being WT-free

$\Rightarrow$ For $t = 0, 1, 2, \ldots$

$$P_0(t) = (1-\alpha)^{2^{t+1}-2} , \quad P_1(t) = 0$$

For each fixed N, we have

$$P_0(r) = \left(1 - \frac{r}{N}\right)^{\frac{2(N-1)}{\log_2 N}}$$

↓

can be measured and evaluated

Chapter 11 : Evolutionary Escape

① Posets and Distributive Lattices

(1) Binary Sequence Space

↳ "0" : unmutated ; "1" : mutated (irreversible)

↳ Genotype := binary string of length n

↳ Wild Type :

$$0 = 00 \cdots 0$$

↳ Escape Type :

$$1 = 11 \cdots 1$$

(2) Partially Ordered Sets (Posets) (there are pairs of elements that cannot be compared)

A poset is a set $\mathcal{E}$ together with a binary relation "$\leq$", which is

- reflexive : $\forall e \in \mathcal{E}$, $e \leq e$
- antisymmetric : $\forall e_1, e_2 \in \mathcal{E}$, if $e_1 \leq e_2$ and $e_2 \leq e_1$, then $e_1 = e_2$
- transitive : $\forall e_1, e_2, e_3 \in \mathcal{E}$, if $e_1 \leq e_2$ and $e_2 \leq e_3$, then $e_1 \leq e_3$

↳ Write $e_1 < e_2$ if $e_1 \leq e_2$ and $e_1 \neq e_2$, $e_1 < e_2$ is called a cover relation if there is no $e' \in \mathcal{E}$ s.t. $e_1 < e' < e_2$.

↳ Hasse Diagram : $G = (\mathcal{E}, E)$ and $e_1 \to e_2 \in E \iff e_1 < e_2$ is a cover relatio

(3) Order Ideal

An order ideal, g, in a poset $\mathcal{E}$ is a subset of $\mathcal{E}$ that is closed downward

i.e. if $e_2 \in g$ and $e_1 \leq e_2$, then $e_1 \in g$.

(4) Distributive Lattice

The set of all order ideals of $\mathcal{E}$ forms a distributive lattice $J(\mathcal{E})$ under inclusion.

↳ $(J(\mathcal{E}), \subseteq)$ is a poset

↳ every pair of order ideals (g_1, g_2) has a unique supremum, $g_1 \cup g_2$, and a unique infimum, $g_1 \cap g_2$.

(5) Genotype Lattice

- Let $\mathcal{E}$ be a set of $n = |\mathcal{E}|$ irreversible genetic events
- Since evolution may not proceed in any random order, the posets $(\mathcal{E}, \leq)$ encode constraints on the order in which the mutations can accumulate.
- The order ideals g of $J(\mathcal{E})$ are the genotype that can evolve subject to order constraints

$\Rightarrow$ $\mathcal{G} = J(\mathcal{E}) =$ Genotype Lattice

(6) The Empty Poset

The empty poset is defined as the set $\mathcal{E} = \{1, 2, \dots, n\}$ with no relations

Then, the genotype lattice $\mathcal{G}$ is simply the hypercube $\underset{\{0,1\}^n}{}$. ⇒ no constraints

(7) Chains

A chain in $\mathcal{G} = (J(\mathcal{E}), \subseteq)$ of length k is a collection of k totally ordered subsets

$$\underbrace{g_1 \subset g_2 \subset \cdots \subset g_k}_{\text{one must be a subset of the next}}$$

The chains in $\mathcal{G}$ are mutational pathways consistent with the poset $\mathcal{E}$.

8) Fitness Landscapes

A fitness landscape is a mapping $f : \mathcal{G} \to \mathbb{R}$

↳ for every genotype $g \in \mathcal{G}$, there is a corresponding fitness

) Evolutionary Escape

We consider a setup where only the escape type has fitness $f > 1$, whereas other genotypes have fitness $f < 1$. In this case, only the escape type has the chance to replicate in the long run, while others will die out eventually with probability 1.

Our question is : under selective pressure, what is the probability that the wild type reaches the escape state before extinction ?

11) Mutational Neighbourhood

The mutational neighbourhood of a genotype $g \in \mathcal{G}$ is the set of genotypes $h \in \mathcal{G}$ that can be reached by mutation :

$$N(g) = \{ h \in \mathcal{G} : g \subset h \}$$

Let $m = |\mathcal{G}| \leq 2^n$ be the number of mutations / genotypes.

Let μ_e be the mutation rate of event $e \in \mathcal{E}$.

Assume that mutations are independent and fix a total order of $\mathcal{G}$, we can write the mutation matrix $\underset{m \times m}{U} = (u_{gh})_{g,h \in \mathcal{G}}$ by

$$u_{gh} = \begin{cases} \prod_{e \in h \setminus g} \mu_e \,, & \text{if } h \in N(g) \\ 0 \,; & \text{otherwise} \end{cases}$$

(2) K-Step Offspring

Let $f : \mathcal{G} \to \mathbb{R}$ be a fitness landscape and set $\underset{m\times m}{F} = \operatorname{diag}(f)$

↳ $(UF)_{g,h}$:= probability of genotype g producing offspring of type h in one step

↳ $(UF)^{k}_{g,h}$:= probability of genotype g producing offspring of type h along any mutational pathway of length k in $\mathcal{G}$

Define a matrix B such that

$$B = UF + (UF)^2 + \cdots + (UF)^n = (I - UF)^{-1} - I$$

where $B = (b_{gh})_{g,h \in \mathcal{G}}$:

$$b_{gh} = \begin{cases} u_{gh}\, f(h)\, P_{gh}(f) & , \text{ if } g \subset h \\ 0 & , \text{ else} \end{cases}$$

where $P_{gh}(f)$ is a polynomial of degree $|h \setminus g| - 1$ on $\mathbb{R}^{\mathcal{G}}$

↳ generating function for all $\underset{\text{pathways}}{\text{chains}}$ from g to h in $\mathcal{G}$

(3) Risk Polynomial

Consider $g = 0$:= wild type and $h = 1$:= escape type.

Write $f(g) = f_g \quad \forall\, g \in \mathcal{G}$.

Then, the risk Polynomial is defined as

$$R(\mathcal{G}, f) := P_{01}(f) = \underbrace{\sum_{0 = g_0 \subset g_1 \subset \cdots \subset g_{k-1} \subset g_k = 1}}_{\text{sum over all chains of length } k} f_{g_1} f_{g_2} \cdots f_{g_{k-1}}$$

(4) Invasion

Let R_g := basic reproductive ratio of an invading pathogen g
= # of offspring a single individual will produce

We are interested in the case : $R_1 > 1$ and $R_g < 1$ for all $g \neq 1$

Define the fitness landscape by :

$$f_g = \frac{R_g}{1 - R_g} = R_g + R_g^2 + R_g^3 + \cdots$$

$$\Leftrightarrow \quad R_g = \frac{f_g}{1 + f_g}$$

For $g \neq 1$, $\quad f_g \approx R_g$

) Multitype Branching Process

Consider a branching process on the type space $\mathcal{G}$ with a Poisson offspring distribution

The probability that a single individual of type g produces k offsprings of type h is

$$p_{gh}^{k} = \frac{(u_{gh} R_g)^k \, e^{-u_{gh} R_g}}{k!} \qquad p_{gh} \sim \text{Poisson}(u_{gh} R_g)$$

Let $\xi_g :=$ probability of escape starting with one individual of type g (risk of escape)

$\Rightarrow \; 1 - \xi_g :=$ probability of extinction

Since for extinction of type g, all lineages of type g must go extinct, We get a recursive formula

$$1 - \xi_g = \prod_{h \geq g} \sum_{k=0}^{\infty} p_{gh}^{k} \, (1 - \xi_h)^k$$

multiply over the mutational neighbourhood of g

chance of g producing k offsprings of h

all k offsprings go extinct

Substituting the Poisson distribution gives:

$$\begin{aligned} 1 - \xi_g &= \prod_{h \geq g} \sum_{k=0}^{\infty} \frac{(u_{gh} R_g)^k}{k!} e^{-u_{gh} R_g} (1 - \xi_h)^k \\ &= \prod_{h \geq g} e^{-u_{gh} R_g} \sum_{k=0}^{\infty} \frac{[u_{gh} R_g (1 - \xi_h)]^k}{k!} \\ &= \prod_{h \geq g} e^{-u_{gh} R_g} \cdot e^{u_{gh} R_g (1 - \xi_h)} \qquad \text{by Taylor series } e^x = \sum_{k=0}^{\infty} \frac{x^k}{k!} \end{aligned}$$

$$\Rightarrow \log(1 - \xi_g) = -\sum_{h \geq g} u_{gh} R_g \xi_h$$

For $g \neq 1$, we have $\xi_g << 1$ and $(R_g)^2 \approx 0$. Thus

$$\begin{aligned} -\xi_g &\cong -R_g \sum_{h \geq g} u_{gh} \xi_h \qquad \text{by } \log(1 - \xi_g) \approx -\xi_g \\ \Rightarrow \quad \xi_g &\approx R_g \Big(\xi_g + \sum_{h > g} u_{gh} \xi_h \Big) \\ \Rightarrow \quad (1 - R_g) \xi_g &\approx R_g \sum_{h > g} u_{gh} \xi_h \\ \Rightarrow \quad \xi_g &\approx f_g \sum_{h > g} u_{gh} \xi_h \qquad \text{by } f_g := \frac{R_g}{1 - R_g} \end{aligned}$$

In particular,

$$\xi_0 \approx f_0 \sum_{h \in \mathcal{G}} u_{0h} \xi_h$$

Solving it yields

$$\xi_0 \approx \xi_1 \cdot f_0 \cdot \prod_{e \in \mathcal{E}} \mu_e \cdot R(\mathcal{G}; f)$$

probability of escape of one wild type

probability of escape of one escape type

fitness of WT

accumulate all necessary mutations

generates all mutational pathways from the wild type to the escape type

By i.i.d. assumption, the probability of escape of N wild-type pathogens is

$$1 - (1 - \xi_0)^N \approx 1 - e^{-N \xi_0}$$

④ The critical Population Size

We define the critical population size as

$$N^* = \frac{1}{\xi_0} \iff 1 - e^{-N^*\xi_0} \approx 1 - \frac{1}{e} \approx \frac{1}{2}$$

↳ If $N >> N^*$, then the escape is almost certain

↳ If $N = N^*$, then the probability of successful intervention is $\frac{1}{e} \approx \frac{1}{2}$

↳ If $N << N^*$, then escape is almost impossible

⇒ Successful treatment depends on the tumour size!

* The more constraints are imposed on $\mathcal{G}$, the larger the critical population size

Chapter 12 Coalescent Theory

① The Coalescent

- In the Wright-Fisher Process, each individual in the new generation chooses a parent cell from the previous generation uniformly at random.
  ↳ there is a certain probability that two or more individuals have one common ancestor
- The coalescent process = thinking the Wright-Fisher Process backward in time
  ↳ there is a certain probability that two or more individuals coalesce
- Coalescent events = branching points in the process, where two or more lineages meet/coalesce
- Coalescent times = the waiting time between j and (j−1) lineages, $T(j)$
  /the waiting time until the next coalescent event
  /branch lengths in the tree
- The probability that j individuals have no common ancestor in the previous generatio

$$j=2: \quad 1-\frac{1}{N}$$

$$j=3: \quad \left(1-\frac{1}{N}\right)\left(1-\frac{2}{N}\right)$$

$$\vdots$$

$$\prod_{i=1}^{j-1}\left(1-\frac{i}{N}\right) = 1-\binom{j}{2}N^{-1}+\underbrace{O(N^{-2})}_{\to 0 \text{ as } N\to\infty}$$

- We measure time in units of N generations.
  Let T(j) be the coalescent time between j and j−1 lineages
  = time in which j individuals have no common ancestor.

$$\mathbb{P}(T(j)>t) = \left(\prod_{i=1}^{j-1}\left(1-\frac{i}{N}\right)\right)^{Nt}$$

$$= \left(1-\binom{j}{2}N^{-1}+O(N^{-2})\right)^{Nt}$$

$$\to \exp\left(-\binom{j}{2}t\right) \qquad \text{as } N\to\infty$$

⇒ only pairwise coalescent events occur in the limit

⇒ $T(j) \sim \text{exponential}\left(\binom{j}{2}\right)$

⇒ The stochastic process that models the coalescent time is called the coalescent

2) Time to the Most Recent Common Ancestor (MRCA)

- Definition : the most recent common ancestor refers to the root of the smallest tree where the leaves are the set of individuals in consideration
- Time to MRCA : for a sample of size n ($\neq$ population size N)

$$T_{MRCA}(n) = \sum_{j=2}^{n} T(j)$$

- Expectation :

$$\mathbb{E}[T_{MRCA}(n)] = \sum_{j=2}^{n} \mathbb{E}[T(j)] \quad \text{by linearity of expectation}$$

$$= \sum_{j=2}^{n} 1/\binom{j}{2} \quad \text{since } T(j) \sim \exp(\tbinom{j}{2})$$

$$= \sum_{j=2}^{n} \frac{2}{j(j-1)}$$

$$= \sum_{j=2}^{n} 2\left(\frac{1}{j-1} - \frac{1}{j}\right)$$

$$= 2\left(1 - \frac{1}{n}\right) \quad \text{by telescoping sum}$$

$\ast$ $\mathbb{E}[T(2)] = 1$: expected time for the last coalescent event is 1 in the unit of N generations

$\ast$ $\lim_{n\to\infty} \mathbb{E}[T_{MRCA}(n)] = 2 = 2 \cdot \mathbb{E}[T(2)]$: twice as long as if there are only 2 individuals ! (50%)

- Variance :

$$\text{Var}[T_{MRCA}(n)] = \sum_{j=2}^{n} \text{Var}(T(j)) \quad \text{since } T(j)\text{'s are non-overlapping}$$

$$= \sum_{j=2}^{n} 1/\binom{j}{2}^2 \quad \text{since } T(j) \sim \exp(\tbinom{j}{2})$$

$$= \sum_{j=2}^{n} \left(\frac{2}{j(j-1)}\right)^2$$

$$= 8\sum_{j=1}^{n} \frac{1}{j^2} + \frac{4}{n^2} - 8\left(1 - \frac{1}{n}\right) - 4$$

$\ast$ $\text{Var}[T(2)] = 1$: variance of the waiting time for the last coalescent event is also 1

$\ast$ $\lim_{n\to\infty} \text{Var}[T_{MRCA}(n)] = \frac{8\pi^2}{6} - 12 \approx 1.16$: the large proportion of this variance is explained by 2-individual coalescent time variance (86%)

(5) Detecting Selection

- Simulation of mutation process

(1) Simulate a tree according to the coalescent process

(2) Superimpose a Poisson process that puts down mutations independently on all branches at rate

$$\frac{\theta}{2} \quad \text{where} \quad \theta = 2Nu \quad \text{(scaled mutation rate)}$$

- Infinite sites model

Assume an infinite number of sites and each mutation to affect a different nucleotide site (unique mutation)

↳ only appropriate for long DNA sequences with uniform mutation rate across sites

- Number of segregating sites : # of sites where not all alleles are identical

Under the infinite sites model,

$$S = \text{total number of mutations}$$

The total branch length is

$$T_{tot}(n) = \sum_{j=2}^{n} j\,T(j)$$

$$\Rightarrow \mathbb{E}[S] = \frac{\theta}{2}\,\mathbb{E}[T_{tot}(n)] \quad \text{by the Poisson process}$$

$$= \frac{\theta}{2} \sum_{j=2}^{n} j \cdot \left(1 \Big/ \binom{j}{2}\right) \quad \text{since } T(j) \sim \exp\left(\binom{j}{2}\right)$$

$$= \theta \underbrace{\sum_{j=2}^{n} \frac{1}{j-1}}_{C_n \,:\, \text{constant that only depends on } n}$$

$$\Longleftrightarrow \theta = C_n^{-1}\,\mathbb{E}[S]$$

- Average Pairwise Nucleotide Distance

$$\boxed{\begin{array}{c}\mathbb{E}[K] = \theta = C_n^{-1}\mathbb{E}[S] \\ \text{under neutral infinite sites model}\end{array}}$$

$$K = \sum_{\substack{i,j \in [n] \\ i \neq j}} \| g_i - g_j \|_0$$

⇒ average pairwise Hamming distance between 2 sequences in the sample

$$\Rightarrow \mathbb{E}[K] = \frac{\theta}{2} \cdot 2\,\underbrace{\mathbb{E}[T(2)]}_{=1} = \frac{\theta}{2} \cdot 2 = \theta$$

- Effects of selections on S and K

(1) S is not sensitive to allele frequency but sensitive to low-frequency deleterious alleles

(2) K is sensitive to allele frequency but not sensitive to low-frequency deleterious alleles

- Tajima's D

$$D = \frac{\hat{K} - C_n^{-1}\hat{S}}{\sqrt{\hat{V}}}$$

- Null hypothesis : there is no selection / the process is neutral
- the distribution of D under the null can be obtained through simulation of the Coalescent process many times
- can calculate the p-value of D using the null distribution and the data

4) Inference under the coalescent

- model parameters θ : mutation rate, population size, selective advantage, ...
- likelihood :

$$L(\theta) = \mathbb{P}(D \mid \theta) = \int \underbrace{P(D \mid T, \theta)}_{\text{statistical phylogenetic tree model}} \underbrace{P(T \mid \theta)}_{\text{coalescent}} \, dT$$

- use MCMC to approximate integral
- model parameters are estimated by MLE or Bayesian inference.

Chapter 15 Tumour Archeology

① Tumor Evolution

- Intra-tumour heterogeneity : there could be multiple clones within a tumour
- Clone: a group of tumour cells that share a highly similar genotype and mutational profile
- Subclone : a group of tumour cells that diverge from an ancestral clone by acquiring additional mutations
- Clone expansion : process in which one genotype with higher fitness expands in frequency in the tumour mass
- Selective sweep : process in which a genotype with a very high fitness emerges and outcompetes all other clones in the tumour (e.g. cancer)
- Driver mutations : mutations that confer a fitness advantage
- Passenger mutations : mutations that have no effect on fitness
- Truncal mutations : ancestral mutations in the trunk of the phylogenetic tree that are shared by all clones
- Subclonal mutations : mutations in a lineage that has diverged from the trunk
- Models of tumour evolution :

  (1) Linear (2) Neutral (3) Branching (4) Punctuated

② Variant-Allele-Frequency-Spectrum

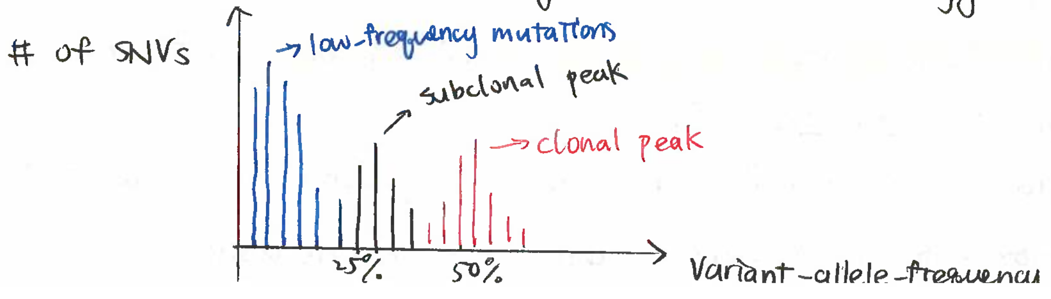
- Binary leaf-labeled tree T : cells are the leaves and mutations occur on branches
- Relation matrix M :

$$M_{ij} = \begin{cases} 1 & \text{if cell } j \text{ is located below mutation } i \text{ in } T \\ 0 & \text{otherwise} \end{cases}$$

- Mutation frequencies :

$$\text{freq}(i) = \sum_j M_{ij}$$

- Data is usually from a mixture of cells .
- Mutations usually occur on a single strand ⇒ heterozygous ⇒ 50% frequencies



) Inference under the neutral model

- Starting with a single cell, the number of tumour cells at time t under exponential growth is

$$N(t) = e^{\lambda\beta t}$$

where λ : cell division rate

β : successful division rate

- Assumptions: Infinite Sites Model

(1) Founding cell has acquired all mutations that give fitness advantage

(2) Subclonal mutations are neutral

- Expected number of new mutations per time interval :

$$\frac{dM(t)}{dt} = \mu \pi \lambda N(t)$$

where μ : mutation rate at cell division

π : ploidy (# of chromosome sets : 2 for human)

- Total number of mutations in interval $[t_0, t]$:

$$M(t) = \mu \pi \lambda \int_{t_0}^{t} N(t)\, dt = \frac{\mu \pi}{\beta}\left(e^{\lambda\beta t} - e^{\lambda\beta t_0}\right)$$

- Mutation age t vs. allele frequency f :

$$f = \frac{1}{\pi N(t)} = \frac{1}{\pi e^{\lambda\beta t}}$$

$$f_{max} = \frac{1}{\pi e^{\lambda\beta t_0}} = \frac{1}{2} \quad \text{for } \pi = 2 \text{ and } t_0 = 0$$

$$\Leftrightarrow \quad e^{\lambda\beta t} = \frac{1}{\pi f}$$

↳ older mutations = higher frequency

right slope of neutral cluste

proportional to

$$\Rightarrow M(f) = \underset{\substack{\downarrow \\ \text{mutation rate per effective cell division}}}{\frac{\mu}{\beta}}\left(\frac{1}{f} - \frac{1}{f_{max}}\right) \quad \text{or} \quad \frac{dM}{df} = -\mu\pi\lambda\frac{1}{f^2}$$

for frequency f, $M(f)$ is the area under the VAF spectrum for all frequencies $> f$.

$\Rightarrow$ there should be a linear relationship b/w $M(f)$ and $\frac{1}{f}$

- Inference for Neutral Evolution :

↳ If the relationship is nonlinear, then the process is not neutral.

↳ If the relationship is linear, then it is inconclusive.

(4) Inference in the presence of selection

- Assume 2 cell populations : host tumour and subclone with different growth rates

$$\lambda_{sub} \geq \lambda_{host}$$

- Selective advantage :

$$s = \frac{\lambda_{sub} - \lambda_{host}}{\lambda_{host}}$$

- Estimating subclone properties from the VAF spectrum

↳ mutation rate μ : estimated from the right slope of the neutral cluster

↳ subclone frequency f_{sub} : estimated from the mean of the subclone peak

↳ number of mutations in the subclone M_{sub} at the time t_1 (when subclone appears): estimated from the area under the VAF curve of the subclone cluster

= the number of mutations acquired between t_0 and t_1

$$\Rightarrow \quad M_{sub} = \mu \Gamma = \mu(2\log(2)\, t_1)$$

where Γ := mean # of sucessful cell divisions b/w t_0 and t_1

$$\Rightarrow \quad t_1 \cong \frac{\hat{M}_{sub}}{2\log(2)\,\hat{\mu}} = \text{age of subclone in terms of doubling}$$

↳ can also estimate the selective advantage :

$$s = \frac{\log\left(\frac{f_{sub}}{1-f_{sub}}\right) + \lambda t_1}{\lambda(t_{end} - t_1)}$$

where t_{end} can be estimated from tumour size

$$2^{t_{end}} = \underbrace{(1 - f_{sub})\, 10^{10}}_{\text{size of tumour}}$$

and $\lambda \triangleq \log(2)$

(5) Confounders of the VAF Spectrum

(1) Contamination with normal cells can shift VAF spectrum to the left

(2) Copy number changes can shift VAF spectrum to the right

(3) Mutation losses destroy connection b/w allele frequency and age