

**Supplement Material for Mathematical Biology**  
**On Models of Population Genetics**

Hong Qian

October 28, 2003

Department of Applied Mathematics  
University of Washington, Seattle

# CHAPTER 1

## BASIC POPULATION DYNAMICS

### 1 Mendelian dynamics and Hardy-Weinberg equilibrium [24]

Let  $P$ ,  $2Q$ , and  $R$  be the number of three genotypes,  $A_1A_1$ ,  $A_1A_2$ ,  $A_2A_2$  made of two alleles  $A_1$  and  $A_2$ , and  $a$ ,  $b$ , and  $c$  be the three respective survival probabilities, and  $r$  the rate of random mating (panmixia). For non-overlapping with separated generations, the next generation

$$P_{n+1} = r(\hat{a}P_n + \hat{b}Q_n)^2 \quad (1)$$

$$Q_{n+1} = r(\hat{a}P_n + \hat{b}Q_n)(\hat{b}Q_n + \hat{c}R_n) \quad (2)$$

$$R_{n+1} = r(\hat{b}Q_n + \hat{c}R_n)^2 \quad (3)$$

We can introduce a set of new parameters called selection coefficients:  $a = \sqrt{r}\hat{a}$ ,  $b = \sqrt{r}\hat{b}$  and  $c = \sqrt{r}\hat{c}$ . Then

$$P_{n+1} = (aP_n + bQ_n)^2 \quad (4)$$

$$Q_{n+1} = (aP_n + bQ_n)(bQ_n + cR_n) \quad (5)$$

$$R_{n+1} = (bQ_n + cR_n)^2 \quad (6)$$

We see that immediately after the first generation,  $P_nR_n = Q_n^2$ . In fact, if  $ac = b^2$  mathematically this is a stable fixed point (equilibrium). This is known as Hardy-Weinberg (HW) equilibrium. However, contrary to Fisher's claim on page 324 of [13], there is no condition on selection coefficients necessary for obtaining HW *proportion* if one allows the growth, and the equilibrium proportions are  $P : 2Q : R = (b - c)^2 : 2(b - a)(b - c) : (b - a)^2$ .

The population size may grow, but one is interested in the population ratio rather than the absolute values. Let  $u = \frac{P}{Q}$  and taking the HW equilibrium into consideration, one has

$$\boxed{u_{n+1} = \frac{au_n^2 + bu_n}{bu_n + c}} \quad (7)$$

Eq. 7 has two equilibrium points:  $u^* = 0$  and  $u^* = \frac{c-b}{a-b}$ .  $u^* = 0$  means  $A_1$  is extinct;  $u^* = \infty$  means  $A_2$  is extinct. These are trivial HW equilibria. The condition for steady-state coexistence of  $A_1$  and  $A_2$  with non-trivial HW equilibrium is  $\frac{b}{c} > 1 > \frac{a}{b}$ . This is known as overdominance.

Finally, we see that because of the HW equilibrium one can also recast Eq. 7 in terms of allelic frequency:  $p = \frac{\sqrt{P}}{\sqrt{P} + \sqrt{R}}$ . Note that in terms of allelic frequency:  $u = \frac{p}{1-p}$ . We thus have

$$p_{n+1} = \frac{ap_n^2 + bp_nq_n}{ap_n^2 + 2bp_nq_n + cq_n^2} \quad (8)$$

in which  $q_n = 1 - p_n$ .

Eq. 7 is the starting point of much of J.B.S. Haldane's work [23] and Eq. 8 is the starting point for R.A. Fisher [13] and S. Wright [51].

## 2 The multiallelic dynamics and its selection potential [8, 51]

Let  $S_{ij}$  be the survival probability of zygote  $A_iA_j$  ( $i, j = 0, 1, \dots, N$ ), which has the frequency according to Hardy-Weinberg proportion  $(2 - \delta_{ij})p_i p_j$  [24]. The mean survival probability, thus, is

$$\bar{w} = \sum_{i=0}^N \sum_{j=0}^N S_{ij} p_i p_j. \quad (9)$$

Following Eq. 8 [51], the change in the frequency for allele  $A_k$  is

$$\Delta p_k = \frac{\sum_{i=0}^N S_{ik} p_i p_k}{\bar{w}} - p_k \quad (10)$$

$$= \frac{\sum_{i=0}^N S_{ik} p_i p_k - p_k \sum_{i=0}^N \sum_{k=0}^N S_{ik} p_i p_k}{\bar{w}} \quad (11)$$

$$= \frac{\sum_{j=0}^N \sum_{i=0}^N S_{ij} p_i p_j (\delta_{jk} - p_k)}{\bar{w}} \quad (12)$$

$$= \frac{\sum_{j=0}^N D_{kj} \sum_{i=0}^N S_{ij} p_i}{\bar{w}} \quad (13)$$

$$= \frac{1}{2} \sum_{j=0}^N D_{kj} \left( \frac{\partial \ln \bar{w}}{\partial p_j} \right)_{p'_i} \quad (14)$$

in which  $p'_i = \{p_i | 0 \leq i \leq N, i \neq j\}$ , and

$$D_{ij} = p_i (\delta_{ij} - p_j). \quad (15)$$

Note that  $\sum_{i=0}^N p_i = 1$  is preserved in the dynamics. The set of independent variables is  $(p_1, p_2, \dots, p_N)$ :

$$\Delta p_k = \frac{1}{2} \sum_{j=1}^N D_{kj} \frac{\partial \ln \bar{w}}{\partial p_j} + \frac{1}{2} D_{k0} \frac{\partial \ln \bar{w}}{\partial p_0}. \quad (16)$$

Since  $D_{ij}$  is a positive definite matrix, let's assume that  $D^{-1} = A^2$  and  $A$  is itself positive definite. Then there exists a convex function  $U(p_0, \dots, p_N)$  as the solution to a Poisson equation:  $\partial^2 U / \partial p_i \partial p_j = A_{ij}$ . Thus  $x_i = \partial U / \partial p_i$  is a one-to-one coordinate transformation, and

$$dx_i = \sum_{j=0}^N A_{ij} dp_j. \quad (17)$$

And finally we have transformed Eq. 14 as a gradient system

$$\Delta x_k = \frac{1}{2} \frac{\partial \ln \bar{w}}{\partial x_k}. \quad (18)$$

The equivalence between Eqs. 14 and 18 reflects a deep geometric nature of the Mendelian dynamics [46]. The dynamics not only has a Lyapunov function  $\ln \bar{w}$ , it actually follows its gradient [52]. This implies both Fisher's theorem that mean fitness always increases and Kimura's theorem that natural selection acts so as to maximize the rate of increase in the (logarithm of) mean fitness of the population.

For two-allele system, we have [21] (Eq. 8)

$$\Delta p = \frac{p(1-p) \{p(S_{11} - S_{10}) - (1-p)(S_{00} - S_{10})\}}{(1-p)^2 S_{00} + 2p(1-p)S_{10} + p^2 S_{11}} \quad (19)$$

*Dominant selection.* The two homozygotes have different fitness values, whereas the fitness of the heterozygote is the same as the fitness of one of the two homozygous genotypes. If the heterozygous fitness is identical to the homozygous genotype with higher fitness, it is called *dominant*,  $S_{00} < S_{10} = S_{11}$ . If the heterozygous fitness is identical to the homozygous genotype with lower fitness, it is called *recessive*,  $S_{00} = S_{10} < S_{11}$ .

*Codominant (genic) selection.* The two homozygotes have different fitness values, whereas the fitness of the heterozygote is the mean of the fitnesses of the two homozygous genotype,  $2S_{10} = S_{00} + S_{11}$ .

*Overdominant selection.* The heterozygote has the highest fitness.  $S_{10} > S_{00}, S_{11}$ . The hallmark of overdominant selection is the coexistence of both alleles with stable equilibrium. It belongs to a class of selection mechanism known as balancing selection.

*Underdominant selection.* The heterozygote has the lowest fitness.  $S_{10} < S_{00}, S_{11}$ . It has a coexistence of both alleles, but the equilibrium is not stable. An allele may be eliminated from

the population even if the fitness of its homozygote is much higher than that of the prevailing homozygote. The hallmark of underdominant selection is that the allele with lower fitness can survive and the an allele with high fitness but low population might still go extinct.

Eq. 19, the basic equation for deterministic population genetics, with two alleles, according to S. Wright, can be written in another equivalent form due to J.B.S. Haldane [22]. Let  $u = \frac{p}{1-p}$ , then we have (Eq. 7)

$$\boxed{u_{n+1} = \frac{S_{11}u_n^2 + S_{10}u_n}{S_{10}u_n + S_{00}}} \quad (20)$$

### 3 Continuous generations: zygote and allelic populations, frequency vs. population dynamics

#### HARDY-WEINBERG DISEQUILIBRIUM IN ZYGOTE POPULATIONS

While the discrete model with separated generations is well established, continuous models are complex and there are many different models. Let  $M_{11}$ ,  $M_{12}$ , and  $M_{22}$  be the number of three genotypes. Assume they die at rates  $d_{ij}$  and their birth  $b_{ij}$ , then

$$\frac{dM_{11}}{dt} = -d_{11}M_{11} + b_{11} \left( M_{11} + \frac{M_{12}}{2} \right)^2 \quad (21)$$

$$\frac{dM_{12}}{dt} = -d_{12}M_{12} + 2b_{12} \left( M_{11} + \frac{M_{12}}{2} \right) \left( M_{22} + \frac{M_{12}}{2} \right) \quad (22)$$

$$\frac{dM_{22}}{dt} = -d_{22}M_{22} + b_{22} \left( M_{22} + \frac{M_{12}}{2} \right)^2 \quad (23)$$

HW equilibrium states that  $M_{11} : M_{12} : M_{22} = p^2 : 2p(1-p) : (1-p)^2$ . Let's introduce a HW disequilibrium parameter  $\theta$

$$\theta \triangleq \frac{4M_{11}M_{22} - M_{12}^2}{(2M_{11} + M_{12})(2M_{22} + M_{12})}, \quad (24)$$

$-1 \leq \theta \leq 1$ . The sufficient and necessary condition for HW equilibrium is  $\theta = 0$ .

The numbers of the alleles 1 and 2 are:

$$N_1 = 2M_{11} + M_{12}, \quad (25)$$

$$N_2 = 2M_{22} + M_{12}. \quad (26)$$

From Eqs. 24, 25, and 26 one can solve  $M_{ij}$

$$M_{11} = \frac{N_1^2 + N_1 N_2 \theta}{2(N_1 + N_2)}, \quad (27)$$

$$M_{12} = \frac{N_1 N_2 (1 - \theta)}{(N_1 + N_2)}, \quad (28)$$

$$M_{22} = \frac{N_2^2 + N_1 N_2 \theta}{2(N_1 + N_2)}. \quad (29)$$

Substituting these into Eqs. 21-23 we have

$$\frac{dN_1}{dt} = -\frac{d_{11}N_1(N_1 + N_2\theta) + d_{12}N_1N_2(1 - \theta)}{N_1 + N_2} + 2b_{11}\left(\frac{N_1}{2}\right)^2 + 2b_{12}\left(\frac{N_1}{2}\right)\left(\frac{N_2}{2}\right) \quad (30)$$

$$\frac{dN_2}{dt} = -\frac{d_{22}N_2(N_2 + N_1\theta) + d_{12}N_1N_2(1 - \theta)}{N_1 + N_2} + 2b_{22}\left(\frac{N_2}{2}\right)^2 + 2b_{12}\left(\frac{N_1}{2}\right)\left(\frac{N_2}{2}\right) \quad (31)$$

$$\begin{aligned} \frac{d\theta}{dt} = & -\frac{(d_{11} + d_{22} - 2d_{12})N_1N_2(1 - \theta)^2}{(N_1 + N_2)^2} + \frac{(b_{11} + b_{22} - 2b_{12})N_1N_2}{2(N_1 + N_2)} \\ & -\frac{(d_{11} - d_{12})N_2 + (d_{22} - d_{12})N_1}{(N_1 + N_2)}\theta(1 - \theta) - \frac{(b_{11} + b_{22})N_1N_2 + b_{12}(N_1^2 + N_2^2)}{2(N_1 + N_2)}\theta \end{aligned} \quad (32)$$

We can see that if  $2d_{12} = d_{11} + d_{22}$  and  $2b_{12} = b_{11} + b_{22}$ , then  $\frac{d\theta}{dt} = 0$  when  $\theta = 0$ . In other words, the population dynamics, Eqs. 21-23 or Eqs. 30-32, preserves the HW proportion.

For large populations, the  $N^2$  terms dominate, so we have the dynamics of two-allele,  $A_1$  and  $A_0$ , system presented by a pair of differential equations irrespective of  $\theta$ :

$$\frac{dN_1}{dt} = r_{12}N_1N_2 + 2r_{11}\left(\frac{N_1^2}{2}\right) \quad (33)$$

$$\frac{dN_2}{dt} = r_{12}N_1N_2 + 2r_{22}\left(\frac{N_2^2}{2}\right) \quad (34)$$

where  $r_{ij} = \frac{b_{ij}}{2}$  are halves of the growth rates for the zygotes  $A_1A_1$ ,  $A_1A_0$ , and  $A_0A_0$ .

We also note that for large  $N_i$ , Eq. 32 becomes

$$\frac{d\theta}{dt} = \frac{(b_{11} + b_{22} - 2b_{12})N_1N_2}{2(N_1 + N_2)} - \frac{(b_{11} + b_{22})N_1N_2 + b_{12}(N_1^2 + N_2^2)}{2(N_1 + N_2)}\theta. \quad (35)$$

Hence, we have a steady Hardy-Weinberg disequilibrium

$$\theta^* = \frac{(b_{11} + b_{22} - 2b_{12})}{(b_{11} + b_{22}) + b_{12}(u + u^{-1})} \quad (36)$$

where  $u = N_1/N_2$ . If all  $b_{ij} = b$ , then only possible steady-state is Hardy-Weinberg equilibrium. This result provides a relationship between the HW disequilibrium and polymorphism, characterized by  $\gamma = \frac{u+u^{-1}}{2}$ .  $\gamma$  being  $\infty$  when there is fixation, and unity when  $A_1$  and  $A_2$  are perfectly balanced:

$$\boxed{\theta^* = \frac{\beta - 1}{\beta + \gamma}} \quad (37)$$

where  $\beta = \frac{b_{11}+b_{22}}{2b_{12}}$ .

### ALLELIC FREQUENCY DYNAMICS: POPULATION SIZE AS A MEASURE OF TIME

If one is interested in the dynamics of allelic frequency,  $f = \frac{N_1}{N_t}$  where  $N_t = N_1 + N_2$  is the size of total population, then from Eqs. 33 and 34 one has

$$\begin{aligned} \frac{df}{dt} &= \frac{d}{dt} \left( \frac{N_1}{N_1 + N_2} \right) \\ &= \frac{N_1 N_2}{2(N_1 + N_2)^2} \left( \frac{\partial \bar{W}}{\partial N_1} \right)_{N_t} \\ &= \frac{f(1-f)}{2} \left( \frac{\partial \bar{W}}{\partial N_1} \right)_{N_t}, \end{aligned}$$

where the average growth rate

$$\bar{W} \triangleq \frac{dN_t}{dt}(N_1, N_t) = (r_{11} - 2r_{12} + r_{11})N_1^2 + 2(r_{12} - r_{22})N_t N_1 + r_{22}N_t^2, \quad (38)$$

$$\left( \frac{d\bar{W}}{dN_1} \right)_{N_t} = 2(r_{11} - r_{12})N_1 + 2(r_{12} - r_{22})N_2. \quad (39)$$

This result should be compared with Eq. 14 from S. Wright. We therefore have a dynamic equation for  $f$  [27]:

$$\frac{df}{d\tau} = f(1-f) \{ (r_{11} - r_{12})f + (r_{12} - r_{22})(1-f) \}. \quad (40)$$

in which an ‘‘effective time’’,  $\tau = \int_0^t N_t(s)ds$ , has been introduced.

In general, for populations  $N_1, \dots, N_n$  with

$$\frac{1}{N_k} \frac{dN_k}{dt} = r_k, \quad (41)$$

we have

$$\boxed{\frac{1}{f_k} \frac{df_k}{dt} = r_k - \sum_{i=1}^n f_i r_i} \quad (42)$$

If all the  $r_k$  are homogeneous functions of  $N$ 's, then it can be written as  $r_k(N) = N_t^\alpha r_k(f_k)$  and Eq. 42 is known as replicator equation [44] with new time variable  $\tau = \int_2^t N_t^\alpha(\xi) d\xi$ . We remark that when population geneticists say “frequency-dependent selection”, what they mean is that  $r_k$  are functions of  $f_j$ , not  $N_j$ . The latter is specifically called density-dependent or population-dependent.

## 4 Coexistence in growing populations [22]

The standard model for coexistence of two competing populations in ecology usually assumes that both populations reach constant steady-states [34]. Population genetists, however, often ask a different question [51]: the coexistence with both populations tending infinity. This question is best answered in terms of the Haldane formalism.

The simple discrete-time Haldane's model in Eq. 20 can be thoroughly analyzed. It is easy to show that the condition to have a nontrivial  $u_n$  when  $n \rightarrow \infty$ , i.e., not zero nor infinity, is when  $S_{10} > S_{00}, S_{11}$ , i.e., the heterozygote being overdominant. To see this, we note that the fixed point for Eq. 20 is

$$u^* = \frac{S_{10} - S_{00}}{S_{10} - S_{11}}, \quad (43)$$

and the approximated linear dynamics near the fixed point is  $u_{n+1} = u_n [1 + r(u_n - u^*)]$  with

$$r = -\frac{(S_{10} - S_{11})^2}{S_{10}^2 - S_{11}S_{00}}. \quad (44)$$

The original proof due to Haldane is instructive. If  $u_n$  is very small approaching to 0, then from Eq. 20 one has  $u_{n+1} = (S_{10}/S_{00})u_n$ . Hence, the condition for  $u_n$  not to become smaller is  $S_{10}/S_{00} > 1$ . Similarly, If  $u_n$  is very large approaching to  $\infty$ , then one have  $u_{n+1} = (S_{11}/S_{10})u_n$ . Hence, the condition for  $u_n$  not to become larger is  $S_{11}/S_{10} < 1$ .

Is this result also held true for two populations with continous time? To see this, let's start with Eqs. 33 and 34:

$$\frac{dN_0}{dt} = N_0 (r_{00}N_0 + r_{10}N_1), \quad (45)$$

$$\frac{dN_1}{dt} = N_1 (r_{10}N_0 + r_{11}N_1), \quad (46)$$

in which  $r$ 's are growth-survival rates. Their relation to the survival probabilities  $S$ 's is  $S = 1+r\Delta t$  where  $\Delta t$  is the step time for the discrete model. Introducing  $u = N_1/N_0$ , we have

$$\frac{du}{dt} = N_0(t)u [(r_{10} - r_{00}) - (r_{10} - r_{11})u]; \quad (47)$$

if one introduces an ‘‘effective time’’  $\tau = \ln N_0(t)$  [note it is different from that in Eq. 40], then Eq. 47 can be written into

$$\frac{du}{d\tau} = \frac{(r_{10} - r_{00})u - (r_{10} - r_{11})u^2}{r_{00} + r_{10}u}. \quad (48)$$

Eq. 48 can have finite positive fixed point (over and underdominance) or not (dominance and codominance). When there is a finite positive fixed point, it can be stable (overdominance) or unstable (underdominance).

Haldane considered an interesting, time-dependent  $r_{00} = f(t)$  problem when  $r_{11} = r_{10} = 1$  [22]. It then can be shown that the condition for nontrivial  $u$  in the limit of  $t \rightarrow \infty$  is

$$\frac{1}{t} \int_0^t f(\xi) d\xi > 1 > \left( \frac{1}{t} \int_0^t \frac{d\xi}{f(\xi)} \right)^{-1} \quad (49)$$

in which the lhs is arithmetic mean and the rhs is harmonic mean. This is an very important result: It states that while for any constant  $f$  dominant selection always leads to extinction, coexistence is possible for fluctuating  $f(t)$ .

## 5 On overdominance with epistasis

Overdominance is particularly interesting since it gives rise to balancing selection. If one considers two independent loci  $A$  and  $B$ , each with two alleles and overdominant heterozygotes, we have the fitness table

|          | $A_1A_1$     | $A_1A_2$ | $A_2A_2$     |
|----------|--------------|----------|--------------|
| $B_1B_1$ | $(1-s)(1-t)$ | $1-t$    | $(1-s)(1-t)$ |
| $B_1B_2$ | $1-s$        | $1$      | $1-s$        |
| $B_2B_2$ | $(1-s)(1-t)$ | $1-t$    | $(1-s)(1-t)$ |

It should be noted that while these two loci are independent, it does not mean the fitness of the genotype are additive - known as epistatic interaction. In fact, if we denote  $x_1, x_2, x_3$  and  $x_4$

as the frequencies of chromosomes  $A_1B_1$ ,  $A_1B_2$ ,  $A_2B_1$  and  $A_2B_2$  respectively, then the dynamic equations for this system is [30, 36]

$$\Delta x_1 = \{x_1(w_1 - \bar{w}) - rS_{14}D\}/\bar{w} \quad (50)$$

$$\Delta x_2 = \{x_2(w_2 - \bar{w}) + rS_{14}D\}/\bar{w} \quad (51)$$

$$\Delta x_3 = \{x_3(w_3 - \bar{w}) + rS_{14}D\}/\bar{w} \quad (52)$$

$$\Delta x_4 = \{x_4(w_4 - \bar{w}) - rS_{14}D\}/\bar{w} \quad (53)$$

in which  $D = x_1x_4 - x_2x_3$ ,  $r$  is the recombination value between the  $A$  and  $B$  loci,  $S_{14} = 1$  is the fitness for double heterozygotes, and  $w_k = (1/2)\partial\bar{w}/\partial x_k$ .

There are three equilibria [4]:

$$x_1^* = \frac{1}{4} \left[ 1 \pm \sqrt{1 - \frac{4r}{st}} \right], \quad x_1^* = \frac{1}{4}, \quad (54)$$

and  $x_2^* = x_3^* = 1/2 - x_1^*$ ,  $x_4^* = x_1^*$ .

The first two equilibria exist only when  $r < st/4$ . Otherwise, the system moves to the third equilibrium. In reality,  $s$  and  $t$  rarely exceed 0.1. Hence  $r$  must be smaller than 0.0025 for these two equilibria to exist.

## 6 Diffusion approximation with genetic drift and mutations [8]

Consider a two-allele system with random sampling, mutation, and constant population  $M$ . The jumping probabilities between the two alleles,  $A \rightleftharpoons A'$ , are  $u$  and  $v$ , respectively, in discrete steps. Let  $N_t^*$  be the number of  $A$  alleles immediately after sampling, which is followed by the mutation. After mutation, the number of  $A$  alleles is  $N_t$ .

The Markov transition probabilities are

$$Pr\{N_t = \ell | N_t^* = m\} = p_{m\ell}, \quad (55)$$

$$Pr\{N_t^* = m | N_{t-1} = n\} = \binom{M}{m} \frac{n^m (M-n)^{M-m}}{M^M}. \quad (56)$$

The generating function for  $p_{m\ell}$  is

$$\sum_{\ell=0}^M p_{m\ell} s^\ell = ((1-u)s + u)^m (vs + 1 - v)^{M-m}. \quad (57)$$

Hence, the conditional expectations

$$\begin{aligned}
E [N_t | N_t^* = m] &= m(1 - u) + (M - m)v \\
E [N_t^2 | N_t^* = m] &= mu(1 - u) + (M - m)v(1 - v) + (m(1 - u) + (M - m)v)^2 \\
E [N_t | N_{t-1} = n] &= n(1 - u) + (M - n)v \\
E [N_t^2 | N_{t-1} = n] &= nu(1 - u) + (M - n)v(1 - v) + M^2v^2 + 2Mv(1 - u - v)n \\
&\quad + (n - n^2/M + n^2)(1 - u - v)^2.
\end{aligned}$$

If one introduces a new random variable  $X_t = \frac{N_t}{M}$ , and let  $x = \frac{n}{M}$ ,  $u = \hat{u}\Delta t$ , and  $v = \hat{v}\Delta t$ , then

$$E \left[ \frac{\Delta X_t}{\Delta t} \middle| X_{t-1} = x \right] = -x\hat{u} + (1 - x)\hat{v} \quad (58)$$

$$\begin{aligned}
E \left[ \frac{(\Delta X_t)^2}{\Delta t} \middle| X_{t-1} = x \right] &= \frac{1}{M\Delta t} \{v + (M - 1)v^2 + (1 - u - 3v - 2(M - 1)v^2 - 2(M - 1)uv)x \\
&\quad - (1 - 2(u + v) + (M - 1)(u + v)^2)x^2\} \\
&= \frac{x(1 - x)}{M\Delta t} \\
&\quad + \frac{1}{M} \{ \hat{v} + \hat{v}^2 - (\hat{u} + 3\hat{v} + 2\hat{v}^2 + 2\hat{u}\hat{v})x - (\hat{u} + \hat{v} - 2)(\hat{u} + \hat{v})x^2 \} \quad (59)
\end{aligned}$$

in which  $\Delta X_t = X_t - X_{t-1}$ .

## 7 Diffusion approximation with genetic drift and selection [8]

Consider a two-allele system with random sampling and selection. The selection is a deterministic process. Let  $N_t^*$  be the number of  $A$  alleles immediately after sampling, which is followed by the selection. After selection, the number of  $A$  alleles is  $N_t$ . We then have

$$\begin{aligned}
E [N_t | N_t^* = My] &= My(1 - y) \{y(r_{11} - r_{10}) - (1 - y)(r_{00} - r_{10})\} \Delta t + My, \\
Pr\{N_t^* = m | N_{t-1} = n\} &= \binom{M}{m} \frac{n^m (M - n)^{M-m}}{M^M}, \\
E [y | N_{t-1} = Mx] &= x, \\
E [y^2 | N_{t-1} = Mx] &= \frac{x(1 - x)}{M} + x^2, \\
E [y^3 | N_{t-1} = Mx] &= \frac{x(x - 1)(2x - 1)}{M^2} + \frac{3x^2(1 - x)}{M} + x^3.
\end{aligned}$$

If one introduces new random variables  $X_t = \frac{N_t}{M}$ ,  $\Delta X_t = X_t - X_{t-1}$ , and let  $x = \frac{n}{M}$ , then

$$E \left[ \frac{\Delta X_t}{\Delta t} \middle| X_{t-1} = x \right] = x(1-x) \{x(r_{11} - r_{10}) - (1-x)(r_{00} - r_{10})\} + O \left( \frac{1}{M} \right), \quad (60)$$

$$E \left[ \frac{(\Delta X_t)^2}{\Delta t} \middle| X_{t-1} = x \right] = \frac{x(1-x)}{M\Delta t} + O \left( \frac{1}{M} \right). \quad (61)$$

## 8 Identical loci problem in finite population [37]

Let's consider a population made up of a constant number of individuals,  $N$ , with  $N_1$  males and  $N_2$  females. The population dynamics, in discrete nonoverlapping generations ( $G$ ), follows Mendelian rule with random mating (panmixia) without mutations. Let  $f_n$  be the probability that the two homologous loci of an individual taken at random from  $G_n$ , the  $n$ th generation, are identical, i.e., they come from the same locus of a common ancestor.

Let's first find out the probability,  $\gamma$ , of the two homologous loci of the individual, taken at random from  $G_n$ , coming from a same individual, male or female, of  $G_{n-2}$ . Note that coming from the same individual can still have nonidentical loci. By enumeration, each individual, male or female, in  $G_{n-1}$  will have  $K \triangleq (2N_1) \times (2N_2)$  possible genotypes. Passing to  $G_n$ , there are total number of  $(2K)^2$  combinations and  $\frac{(2N_1+2N_2) \times (2N_1+2N_2+1)}{2}$  possible genotypes. Among them the number of combinations with both loci from a particular male of  $G_{n-2}$  is  $2K$ , and from a same male of  $G_{n-2}$  is  $2N_1 \times 2K$ . Similarly from a same female of  $G_{n-2}$  is  $2N_2 \times 2K$ . Hence,

$$\gamma = \frac{(2N_1)(2K) + (2N_2)(2K)}{(2K)^2} = \frac{N_1 + N_2}{4N_1N_2}. \quad (62)$$

$1/\gamma$  is the harmonic mean of  $2N_1$  and  $2N_2$ .

Of course, there is probability  $f_{n-2}$  that an individual in  $G_{n-2}$  has identical homologous loci. In this case, the probability for the grandchild to have identical loci is

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

The probability that two homologous loci taken from two different individuals of  $G_{n-2}$  are identical is  $f_{n-1}$ . Hence,

$$f_n = \gamma \left( \frac{1 + f_{n-2}}{2} \right) + (1 - \gamma) f_{n-1}. \quad (63)$$

Eq. 63 can be solved.

$$f_n = 1 - a\lambda_1^n - b\lambda_2^n, \quad (64)$$

where

$$\lambda_{1,2} = \frac{1 - \gamma \pm \sqrt{1 + \gamma^2}}{2}.$$

For large  $N$  the  $\gamma$  is small. Hence  $\lambda_1 \approx 1 - \gamma/2$   $\lambda_2 \approx -\gamma/2$ . If  $f_0 = f_1 = 0$ , when we have  $a = 1$ ,  $b = 0$ , and

$$f_n \approx 1 - \left(1 - \frac{\gamma}{2}\right)^n. \quad (65)$$

So when generation  $n = 1/\gamma$ , the identical loci problem becomes significant. This is related to the concept of effective population size  $N_e$ .

## 9 On the number of distinct individuals

If we have  $N$  distinct individual in generation zero, and by random Bernoulli sampling with constant population  $N$ , what is the number of distinct individual,  $M_k$ , in generation  $k$ ? This is a Markov process in terms of  $(k_1, k_2, \dots, k_N)$  as function of generations. Clearly the number of distinct individuals decreases with generations and eventually becomes one.

In terms of the  $M_k$ , we have

$$Pr\{M_1 = \ell\} = \binom{N}{\ell} \left(\frac{\ell}{N}\right)^N - \binom{N}{\ell-1} \left(\frac{\ell-1}{N}\right)^N, \quad (66)$$

The expectation of  $M_1$

$$E[M_1] = \sum_{k=1}^N \left[1 - \binom{N}{k} \left(\frac{k}{N}\right)^N\right]. \quad (67)$$

In the backward direction, the probability of two given current individuals coming from a same ancestor is

$$\frac{N!}{k_1! \cdots k_N!} \frac{1}{N^N}, \quad (68)$$

and picking one in  $k_\ell$  out if  $N$ :

$$\frac{N!}{k_\ell!(N-k_\ell)!} \left(\frac{1}{N}\right)^{k_\ell} \left(\frac{N-1}{N}\right)^{N-k_\ell} \quad (69)$$

$$Pr\{\} = \sum_{k_\ell} \frac{k_\ell(k_\ell-1)}{N(N-1)} \frac{N!}{k_\ell!(N-k_\ell)!} \left(\frac{1}{N}\right)^{k_\ell} \left(\frac{N-1}{N}\right)^{N-k_\ell} = \frac{1}{N^2}. \quad (70)$$

Therefore, from any same ancestor is  $1/N$ .

## 10 Natural selection versus neutral mutations

Neo-Darwinism advocates that natural selection is the most important mechanism in evolution, more specifically, in changing gene frequencies and creating new species (p. 407 [40]). It is of course accepted that mutation is the primary source of variation, but its effect on evolution is small.

We shall put the above statement in scrutiny.

First, while there is no doubt that both gene frequency change and species creation are important components of evolution, is there a reason to believe that both processes share a single dominant mechanism?

Second, how to quantitatively define “importance” for a mechanism, since clearly there would be no evolution without mutation? The answer to this question is based on a kinetic argument: Is the selection or mutation the “rate limiting step” in evolutionary change?

However, one could define the “importance” differently. The neo-Darwinism suggests that mutations recur with a high frequency and there is a sufficient amount of genetic variability to respond to any selection. In other words, the direction of the evolution is dictated by selection, and mutation is random. So in this sense, clearly the selection is more important. In fact, one could say that the most important mechanism in evolution is neither selection nor mutation, but environmental change which determines the direction of evolution!

This last point has an interesting consequence. In molecular evolution, it is known that the hemoglobins in living fossils have “evolved” nearly at the same rate as those of primates (p. 410 [40]). Clearly the former has a slowly changing environment. However, is the word “evolve” used on protein and DNA sequences and used for population genetics really the same? The answer to this question is in “molecular evolution versus phenotypic evolution” (p. 414 [40]).

The neutral (mutation) theory suggests that most sequence substitutions are caused by random fixation of neutral or nearly neutral mutations.

Finally, the above debate clearly is based on the assumption that the mutation is the sole source of changing genetic sequences. But this needs not to be true since other dynamics processes discovered in the past several decades can also be a source of sequence variability.

# CHAPTER 2

## STOCHASTIC DYNAMICS OF GENETIC EVOLUTION WITH FLUCTUATING ENVIRONMENT

### 11 The mechanism of polymorphism

Balancing selections according to neo-Darwinism and mutation-selection balance according to neutral theory are two competing mechanisms for maintaining polymorphism. Both these two theories are based on the assumption that the polymorphism is a stationary distribution. An alternative mechanism is that the dynamical system is not in its stationary state; rather due to fluctuating environment, it never reaches a steady-state but in constant transient relaxation. We call this equilibrium versus kinetic views of polymorphism.

In terms of the diffusion model for genetic drift and mutation with selection, the evolution is a stochastic process on a “selection” landscape; and it can be absorbed at various places due to genetic fixation. However, if the landscape keep changing, the likely hood of fixation is greatly reduced - this is our environment-kinetic view of polymorphism.

### 12 A toy model of haploid dynamics

Even though diploid population genetics is our main interest, certain insights can be obtained from a study of a simple model for haploids [19]. Let’s consider a pair of alleles  $A_0$  and  $A_1$  for a single locus:

$$\frac{dN_0}{dt} = r_0(t)N_0 \tag{71}$$

$$\frac{dN_1}{dt} = r_1(t)N_1 \tag{72}$$

where the growth-survival rates  $r_0$  and  $r_1$  are both time-varying. If one introduces the relative allele frequency as the variable  $y = \frac{N_1}{N_0}$ , then we have

$$\frac{dy}{dt} = r(t)y, \tag{73}$$

where  $r(t) = r_1(t) - r_0(t)$ . The solution is

$$y(t) = y(0)e^{\int_0^t r(s)ds}. \quad (74)$$

With stochastically fluctuating  $r(t)$ , let's assume it is a stationary Ornstein-Uhlenbeck process:

$$\frac{dr}{dt} = -a(r - \bar{r}) + b\xi(t). \quad (75)$$

The probability distribution density for random variable  $R_t = \int_0^t r(s)ds$  is

$$f_{R_t}(x) = \frac{1}{\sqrt{2\pi\sigma^2t}} \exp \left\{ -\frac{(x - \bar{r}t)^2}{2\sigma^2t} \right\}, \quad (76)$$

in which  $\sigma^2 = (b/a)^2$ . Therefore,  $y(t)$  follows a lognormal distribution:

$$f_{y_t}(y) = \frac{1}{\sqrt{2\pi\sigma^2ty}} \exp \left[ -\frac{\left( \ln \frac{y}{y(0)} - \bar{r}t \right)^2}{2\sigma^2t} \right]. \quad (77)$$

Finally, if we introducing  $x = \frac{y}{1+y}$  as the frequency of allele  $A_1$ , we have its pdf, known as  $S_B$  distribution:

$$f_{x_t}(x) = \frac{1}{\sqrt{2\pi\sigma^2tx(1-x)}} \exp \left[ -\frac{\left( \ln \frac{x(1-x(0))}{(1-x)x(0)} - \bar{r}t \right)^2}{2\sigma^2t} \right] \quad (78)$$

### 13 The Mendelian dynamics of diploid alleles

The basic elements of the model are as follows. Let's consider a pair of alleles  $A_0$  and  $A_1$  for a single locus. The population dynamics for the two alleles is

$$\frac{dN_0}{dt} = r_{00}N_0^2 + r_{10}N_0N_1 \quad (79)$$

$$\frac{dN_1}{dt} = r_{10}N_0N_1 + r_{11}N_1^2 \quad (80)$$

in which the growth-survival rates  $r_{ij}$  are functions of  $X_0$  and  $X_1$ , the activities of enzymes as the gene products of  $A_0$  and  $A_1$ , respectively:

$$r_{11} = \phi(X_1), \quad r_{00} = \phi(X_0), \quad r_{10} = \phi \left( \frac{X_1 + X_0}{2} \right). \quad (81)$$

The enzyme activities, however, are fluctuating stochastic processes,  $X_i(t)$ , due to varying temporal environment. If we assume that both  $X_1$  and  $X_2$  are stationary with small variations fluctuating near 1, then we have

$$\phi(X(t)) = 1 + c_1 Y(t) + c_2 Y^2(t) + o(Y^3) \quad (82)$$

in which  $Y(t) = X(t) - 1$ ,  $c_1 = \phi'(1)$ , and  $c_2 = \frac{1}{2}\phi''(1)$ . One can further assume  $Y_i(t)$  be a continuous-time Markov process with continuous trajectory; Hence it is a two-dimensional diffusion process and can be characterized by the stochastic differential equation

$$\frac{dY_i}{dt} = b_i(Y_0, Y_1) + a_{i1}(Y_0, Y_1)\xi_1(t) + a_{i2}(Y_0, Y_1)\xi_2(t) \quad (i = 0, 1) \quad (83)$$

where  $\xi_i(t)$  are independent white noises (Wiener processes). Substituting Eq. 82 into Eqs. 79 and 80, we have

$$\frac{dN_0}{dt} = (N_0^2 + N_0N_1) + c_1 \left\{ Y_0N_0^2 + \frac{Y_0 + Y_1}{2}N_0N_1 \right\} + c_2 \left\{ Y_0^2N_0^2 + \frac{(Y_0 + Y_1)^2}{4}N_0N_1 \right\} \quad (84)$$

$$\frac{dN_1}{dt} = (N_0N_1 + N_1^2) + c_1 \left\{ Y_1N_1^2 + \frac{Y_0 + Y_1}{2}N_0N_1 \right\} + c_2 \left\{ Y_1^2N_1^2 + \frac{(Y_0 + Y_1)^2}{4}N_0N_1 \right\} \quad (85)$$

Eqs. 83-85 complete the mathematical model.

We now derive the dynamics equation for the frequency of the  $A_1$  allele,  $p = \frac{N_1}{N_0 + N_1}$ .

$$\begin{aligned} \frac{dp}{dt} &= \frac{N_0 \frac{dN_1}{dt} - N_1 \frac{dN_0}{dt}}{(N_0 + N_1)^2} \\ &= \frac{(N_0 + N_1)p(1-p)}{4} \left\{ 2c_1(Y_1 - Y_0) + c_2 [(1 + 2p)Y_1^2 + (2p - 3)Y_0^2 + 2(1 - 2p)Y_0Y_1] \right\}. \end{aligned}$$

Introducing a new time variable  $\tau = \ln(N_0 + N_1)$ , we have

$$\begin{aligned} \frac{dp}{d\tau} &= p(1-p) \left\{ \frac{1}{2}c_1(Y_1 - Y_0) + \frac{c_2}{4} [(1 + 2p)Y_1^2 + (2p - 3)Y_0^2 + 2(1 - 2p)Y_0Y_1] \right. \\ &\quad \left. - \frac{c_1^2}{2} [pY_1^2 - (1-p)Y_0^2 + (1-2p)Y_0Y_1] \right\} \quad (86) \end{aligned}$$

where we have made use of the fact of small fluctuations  $Y_0, Y_1 \ll 1$ .

In order to reach the equation 6 of [17], we assume that the dynamics of  $p(t)$  is much slower than the fluctuation of  $Y(t)$ . Hence one is able to approximate  $E[g(p)h(Y)] \approx g(p)E[h(Y)]$  for arbitrary functions  $g(\cdot)$  and  $h(\cdot)$ . Then we have Eq. 86 becoming

$$\frac{d}{d\tau}\bar{p}(t) = \bar{p}(1-\bar{p}) \left\{ \frac{c_1}{2}(\mu_1 - \mu_0) + \frac{c_2 - c_1^2}{2}\sigma^2(2\bar{p} - 1)(1 - \rho) \right\} \quad (87)$$

in which  $\bar{p} = E[p]$ .

## DYNAMICS OF RELATIVE FREQUENCY: CONTINUOUS MODEL

We now carry out an alternative analysis of the model following the work of Haldane and Jayakar [23] who focused on  $u = \frac{N_1}{N_0}$  rather than the frequency  $p$ . This analysis is both conceptually and technically simpler, and one needs not to make the rapid fluctuation approximations. The dynamic equation for  $u$  is

$$\frac{du}{d\tau} = \frac{(r_{11} - r_{10})u^2 - (r_{00} - r_{10})u}{r_{10}u + r_{00}} \quad (88)$$

where the effective time  $\tau = \ln N_0$ .

It is interesting to note that if instead one introduces an alternative effective time  $\hat{\tau} = \int_0^t (N_0(s) + N_1(s)) ds$ , then

$$\frac{du}{d\hat{\tau}} = \frac{(r_{11} - r_{10})u^2 - (r_{00} - r_{10})u}{u + 1}.$$

This equation when we assume the growth of a heterozygote being exactly the intermediate between that of the two associated homozygotes,  $r_{11} - r_{10} = r_{10} - r_{00} = r$ , yields

$$\frac{du}{d\hat{\tau}} = ru, \quad (89)$$

which is exactly the same equation for Haploid dynamics.

Now if  $r$ 's are functions of  $X$ 's, which themselves are fluctuating with time, all the  $r$ 's are functions of time. <sup>1</sup> Following [23] we have the condition for polymorphic alleles, i.e.,  $0 < u < \infty$ :

$$\frac{1}{t} \int_0^t \frac{r_{11}(s)}{r_{10}(s)} ds < 1 < \frac{1}{t} \int_0^t \frac{r_{10}(s)}{r_{00}(s)} ds \quad (90)$$

If  $X(t)$ 's are stationary stochastic processes, then according to ergodic theorem Eq. 90 can be rewritten as

$$\boxed{E \left[ \frac{r_{11}}{r_{10}} \right] < 1 < E \left[ \frac{r_{10}}{r_{00}} \right]} \quad (91)$$

This is a more compact result than that in [17], and so far it is exact. Substituting  $\phi(\cdot)$  and  $X_i(t)$  into Eq. 91, we have

$$\frac{c_2}{c_1} - c_1 < \frac{\mu_1 - \mu_0}{\sigma^2(1 - \rho)} < -\frac{c_2}{c_1} \quad (92)$$

---

<sup>1</sup>There is an issue of which time should be used for the  $X$ :  $t$ ,  $\tau$ , or  $\hat{\tau}$ ? For standard population genetic theory, this question can be circumvented by assuming the generations are separated, and the dynamics is represented by difference equations rather than differential equations.

which is slightly different from what Gillespie's equation 13 on page 813 of [17]. To verify (t)his result, we can substitute  $\phi(\cdot)$  and  $X_i(t)$  into Eq. 88 and assuming  $Y_0, Y_1 \ll 1$ ,

$$\begin{aligned} \frac{du}{d\tau} = & \frac{c_1 u}{2} (Y_1 - Y_0) + \frac{c_2 u}{4(1+u)} \left\{ (1+3u)Y_1^2 + 2(1-u)Y_0Y_1 - (3+u)Y_0^2 \right\} \\ & - \frac{c_1^2 u}{4(1+u)} \left\{ uY_1^2 + 2Y_0Y_1 - (2+u)Y_0^2 \right\} \end{aligned} \quad (93)$$

and after averaging with rapid fluctuation approximation

$$\frac{d}{d\tau} \bar{u} = \frac{c_1 \bar{u}}{2} (\mu_1 - \mu_0) - \frac{c_2 \bar{u} \sigma^2}{2(1+\bar{u})} (1-\bar{u})(1-\rho) + \frac{c_1^2 \bar{u} \sigma^2}{2(1+\bar{u})} (1-\rho). \quad (94)$$

If now we carry out the analysis following [23] we obtain

$$\frac{c_2}{c_1} - c_1 < \frac{\mu_1 - \mu_0}{\sigma^2(1-\rho)} < -\frac{c_2}{c_1} \quad (95)$$

which is identical to Eq. 92 and agrees with a necessary condition

$$\left| \frac{\mu_1 - \mu_0}{\sigma^2(1-\rho)} \right| < c_1 - \frac{c_2}{c_1}, \quad (96)$$

and a sufficient condition

$$\left| \frac{\mu_1 - \mu_0}{\sigma^2(1-\rho)} \right| < -\frac{c_2}{c_1}. \quad (97)$$

## DYNAMICS OF RELATIVE FREQUENCY: DISCRETE MODEL

Following Haldane and Jayakar, the continuous model gives a condition for time-varying selection coefficients, Eq. 90. The discrete model, on the other hand, gives the condition

$$\left( \prod_{k=1}^n \frac{r_{11}(k)}{r_{10}(k)} \right)^{1/n} < 1 < \left( \prod_{k=1}^n \frac{r_{10}(k)}{r_{00}(k)} \right)^{1/n} \quad (98)$$

which can be rewritten, in the case of stochastic stationary  $r$ 's according to ergodic theorem

$$\exp \left\{ E \left[ \ln \left( \frac{r_{11}}{r_{10}} \right) \right] \right\} < 1 < \exp \left\{ E \left[ \ln \left( \frac{r_{10}}{r_{00}} \right) \right] \right\} \quad (99)$$

which can be further simplified into

$$E [\ln r_{11}] , E [\ln r_{00}] < E [\ln r_{10}] \quad (100)$$

If  $r_{11} = 1 + c_1 Y_1 + c_2 Y_1^2$ , we have for  $Y_1 \ll 1$

$$E[\ln r_{11}] = c_1 \mu_1 + c_2 \sigma^2 - \frac{1}{2} c_1^2 \sigma^2, \quad (101)$$

and for  $r_{10} = 1 + c_1 \frac{Y_1 + Y_0}{2} + c_2 \left(\frac{Y_1 + Y_0}{2}\right)^2$ , we have

$$E[\ln r_{10}] = c_1 \frac{\mu_1 + \mu_0}{2} + c_2 \frac{\sigma^2(1 + \rho)}{2} - c_1^2 \frac{\sigma^2(1 + \rho)}{4}. \quad (102)$$

Therefore, Eq. 100 yields

$$\left| \frac{\mu_1 - \mu_0}{\sigma^2(1 - \rho)} \right| < \frac{c_1}{2} - \frac{c_2}{c_1}. \quad (103)$$

## 14 Diffusion with singular boundary

### STOCHASTIC STABILITY OF DIFFUSION

The stochastic stability result of H. Kushner [35] is about the convergence of a stochastic process to a singular point, which could be a boundary, with probability 1. Taking the simple example on page 55 of [35]

$$dx = axdt + \sigma x dz, \quad (104)$$

by Ito integration the solution to this stochastic differential equation is

$$x(t) = x(0) \exp \left\{ \left( a - \frac{\sigma^2}{2} + \frac{z_t}{t} \right) t \right\}. \quad (105)$$

$x(t)$  converges to 0 as  $t \rightarrow \infty$  if  $2a < \sigma^2$ . This result can also be seen from solving the stationary probability density  $f(x)$  for  $x$  according to Kolmogorov forward equation:

$$\frac{d^2}{dx^2} \frac{\sigma^2 x^2}{2} f(x) - \frac{d}{dx} ax f(x) = 0. \quad (106)$$

The general solution to (106) is

$$f(x) = cx^{\frac{2a}{\sigma^2} - 2} - \frac{2J}{\sigma^2 - 2a} x^{-1}. \quad (107)$$

There are two constants of integration,  $c$  and  $J$ . If  $J \neq 0$ ,  $f(x)$  is non-integrable under any condition. For the diffusion process defined by stochastic differential equation (104) on  $\mathbb{R}^1$ , however, the boundary condition is such that  $J = 0$ . Therefore,  $f(x)$  is integrable at 0 if  $\frac{2a}{\sigma^2} > 1$ . When  $f(x)$

is not integrable at 0, i.e.,  $2a < \sigma^2$ , the the time-dependent solution to the Kolmogorov forward equation tends to a  $\delta$  function at  $x = 0$ . This is the stochastic stability; in population genetic terms, the true fixation occurs.

If  $1 < \frac{2a}{\sigma^2} < 2$ , then  $f(x)$  is integrable at  $x = 0$  but “piling up” there. This is known in population genetics as quasi fixation according to Kimura.

If  $\frac{2a}{\sigma^2} > 2$ , then  $f(x)$  is peaked away from  $x = 0$ , This corresponds to coexistence, i.e., polymorphism. This last condition can also be obtained from a slightly different treatment. Consider the equation for the stationary distribution  $f(x)$  in general:

$$\frac{d^2}{dx^2}a(x)f(x) - \frac{d}{dx}b(x)f(x) = 0. \quad (108)$$

Then

$$a(x)\frac{d^2 f}{dx^2} + \left(2\frac{da(x)}{dx} - b(x)\right)\frac{df}{dx} + \left(\frac{d^2 a(x)}{dx^2} - \frac{db(x)}{dx}\right)f = 0 \quad (109)$$

If there is a maximum of  $f$  at  $x^*$ , then

$$\frac{df(x^*)}{dx} = 0, \quad \frac{df^2(x^*)}{dx^2} < 0. \quad (110)$$

Therefore,

$$\frac{d^2 a(x)}{dx^2} - \frac{db(x)}{dx} = \frac{a(x)}{f(x)} \frac{d^2 f}{dx^2} < 0. \quad (111)$$

at  $x = x^*$ . For the example in Eq. 106,  $a(x) = \frac{\sigma^2 x^2}{2}$  and  $b(x) = ax$ . Hence one derives  $\sigma^2 - a < 0$ .

## BOUNDARIES OF SINGULAR DIFFUSION

In [38] P. Mandl introduced  $m(x)$  and  $p(x)$ . Their relations to  $a(x)$  and  $b(x)$  are defined on page 13 of [38]:

$$B(x) = \int^x b(y)a(y)^{-1}dy \quad (112)$$

$$m(x) = \int^x a(y)^{-1}e^{B(y)}dy \quad (113)$$

$$p(x) = \int^x e^{-B(y)}dy \quad (114)$$

The general solution to Eq. 108 is

$$f(x) = \left\{ c - J \int^x e^{-B(y)}dy \right\} \frac{e^{B(x)}}{a(x)}. \quad (115)$$

in which one of the two constants of integration,  $J$ , has the meaning of boundary flux. The equation adjoint to Eq. 108,

$$a(x) \frac{d^2}{dx^2} g(x) + \frac{d}{dx} g(x) = -1 \quad (116)$$

gives the mean time to stay inside a finite domain, starting from  $x$ ; or the mean time to arrive at  $x$  starting from a point inside the domain. Its general solution is

$$g(x) = c_1 + c_2 \int^x e^{-B(y)} dy \int^y \frac{e^{B(z)}}{a(z)} dz. \quad (117)$$

Mandl further introduces two functions  $u^1(x)$  and  $v^1(x)$  defined on page 24 of [38]:

$$u^1(x) = \int^x m(y) dp(y) = \int^x e^{-B(y)} dx \int^y a(z)^{-1} e^{B(z)} dz, \quad (118)$$

$$v^1(x) = \int^x p(s) dm(s) = \int^x a(y)^{-1} e^{B(y)} dy \int^y e^{-B(z)} dz. \quad (119)$$

We see that  $u^1(x)$  is the mean time to arrive at  $x$ , it also gives the sojourn time at  $x$ .  $v^1(x)$  is the integrability of the density function at  $x$  when  $J \neq 0$ . If  $J = 0$  which is the case for stochastic differential equation (104), one uses instead the integrability condition

$$m(x) = \int^x a(y)^{-1} e^{B(y)} dy. \quad (120)$$

This is the one relevant to the result of H. Kushner. It is also the condition for the existence of a weakly converging asymptotic probability distribution as  $t \rightarrow \infty$  ([38], page 90).

According to Mandl ([38], page 24), a boundary is called accessible if  $u^1(0) < \infty$  and inaccessible if  $u^1(0) = \infty$ . For an accessible boundary, if  $v^1(0) < \infty$  it is *regular* and if  $v^1(0) = \infty$  it is an *exit boundary*. An exit boundary attracts trajectories with probability 1, but on average the trajectories takes infinite long time to reach it. For an inaccessible boundary, if  $v^1(0) < \infty$  it is an *entrance boundary* and if  $v^1(0) = \infty$  it is a *natural boundary*. An entrance boundary has infinite mean sojourn time, but it takes infinite long time to get there and it does not attract trajectories with probability 1. A natural boundary has infinite mean sojourn time, takes infinite long time to get there, and attract with probability 1.

For the example in Eq. 106, we have

$$B(x) = \frac{2a}{\sigma^2} \ln x \quad (121)$$

$$m(x) = \frac{2}{2a - \sigma^2} \left( x^{\frac{2a}{\sigma^2} - 1} - 1 \right) \quad (122)$$

$$p(x) = \frac{\sigma^2}{\sigma^2 - 2a} \left( x^{-\frac{2a}{\sigma^2} + 1} - 1 \right) \quad (123)$$

$$u^1(x) = \frac{2}{(2a - \sigma^2)^2} \left[ (2a - \sigma^2) \ln x + \sigma^2 x^{1 - 2a/\sigma^2} \right] \quad (124)$$

$$v^1(x) = \frac{2}{(2a - \sigma^2)^2} \left[ (\sigma^2 - 2a) \ln x + \sigma^2 x^{2a/\sigma^2 - 1} \right] \quad (125)$$

Therefore, if  $J \neq 0$ , the singularity at  $x = 0$  is classified as natural boundary. If  $J = 0$ , then it is again a natural boundary when  $2a < \sigma^2$ ; but it is an entrance boundary when  $2a > \sigma^2$ .

## IMPLICATIONS TO GENETIC POLYMORPHISM

Applying the above results to Gillespie's 1974 model [16], we have the following conclusions. The condition in the equation (1) on page 148 of [16] is based on the integrability of  $f(x)$  at  $x = 0$ :  $m(0) < \infty$  in Eq. 120. When this does not hold, there is true fixation. However, even when  $m(0)$  is finite, the  $\beta$ -distribution can still have quasi-fixation if

$$\Delta\Gamma + \frac{\sigma^2}{2} < \frac{2}{\alpha}. \quad (126)$$

Under this condition, the  $\beta$ -distribution piles up at both  $x = 0$  and  $x = 1$ . Therefore, we have

$$|\Delta\Gamma| > \frac{\sigma^2}{2} - \frac{1}{\alpha} \quad \text{fixation} \quad (127)$$

$$\frac{\sigma^2}{2} - \frac{1}{\alpha} > |\Delta\Gamma| > \frac{\sigma^2}{2} - \frac{2}{\alpha} \quad \text{quasi fixation} \quad (128)$$

$$|\Delta\Gamma| < \frac{\sigma^2}{2} - \frac{2}{\alpha} \quad \text{polymorphism.} \quad (129)$$

The above equations give the probability of fixation. However, irrespective of the parameters, the mean time to fixation is always  $\infty$ .

Applying the same results to Gillespie's 1976 S2 model [17], we have

$$\frac{|\Delta M|}{\sigma^2(1-\rho)} > \frac{c_1}{2} - \frac{c_2}{c_1} \quad \text{fixation} \quad (130)$$

$$\frac{c_1}{2} - \frac{c_2}{c_1} > \frac{|\Delta M|}{\sigma^2(1-\rho)} > -\frac{c_2}{c_1} \quad \text{quasi fixation} \quad (131)$$

$$\frac{|\Delta M|}{\sigma^2(1-\rho)} < -\frac{c_2}{c_1} \quad \text{polymorphism.} \quad (132)$$

We notice that Eq. 132 is the same as Eq. 97.

## References

- [1] For a brief history of classical evolutionary genetics, see <http://instruct.uwo.ca/zoology/441a/hist.html>.
- [2] Ayala, F.J. and Kiger, J.A. (1984) *Modern Genetics*, 2nd Ed., Benjamin-Cummings, Menlo Park, CA.
- [3] Babajide, A., Hofacker, I.L., Sippl, M., and Stadler, P.F. (1997) Neutral networks in protein space: a computational study based on knowledge-based potentials of mean force. *Fold. Design*, **7**, 261-269.
- [4] Bodmer, W.F. and Felsenstein, J. (1967) Linage and selection: theoretical analysis of the deterministic two locus random mating model. *Genetics*, **57**, 237-265.
- [5] Bornberg-Bauer, E. and Chan, H.S. (1999) Modeling evolutionary landscapes: mutational stability, topology, and superfunnels in sequence space. *Proc. Natl. Acad. Sci. USA*, **96**, 10689-10694.
- [6] Chakraborty, R. and Nei, M. (1976) Bottleneck effects on average heterozygosity and genetic distance with the stepwise mutation model. *Evol.* **31**, 347-356.
- [7] Chan, H.S. and Bornberg-Bauer, E. (2002) Perspectives on protein evolution from simple exact models. *Appl. Bioinformatics*, **1**, 121-144.
- [8] Ethier, S.N. and Kurtz, T.G. (1986) *Markov Processes: Characterization and Convergence*. John Wiley & Sons, New York.
- [9] Ewens, W.J. (1969) *Population Genetics*, Methuen & Co. Ltd, London.
- [10] Feller, W. (1952) The parabolic differential equation and the associated semi-groups of transformations. *Ann. Math.* **55**, 468-519.
- [11] Feller, W. (1954) Diffusion processes in one dimension. *Tran. Am. Math. Soc.* **77**, 1-31.
- [12] Felsenstein, J. (2003) *Theoretical Evolutionary Genetics*, Univ. of Wash., Seattle.

- [13] Fisher, R.A. (1922) On the Dominance Ratio. *Proc. Roy. Soc. Edinburgh*. **42**, 321-341.
- [14] Fontana, W. and Schuster, P. (1998) Continuity in evolution: on the nature of transitions. *Science*, **280**, 1451-1455.
- [15] Fu, Y.-X. and Li, W.-H. (1999) Coalescing into the 21st century: an overview and prospectis of coalescent theory. *Theoret. Popul. Biol.* **56**, 1-10.
- [16] Gillespie, J.H. (1974) Polymorphism in patchy environments. *Amer. Natur.* **108**, 145-151.
- [17] Gillespie, J.H. (1976) A general model to account for enzyme variation in natural popula- tions. II. Characterization of the fitness functions. *Am. Natur.*, **110**, 809-821.
- [18] Gillespie, J.H. (1977) A general model to account for enzyme variation in natural popula- tions. III. Multiple alleles. *Evolution*, **31**, 85-90.
- [19] Gillespie, J.H. (1991) *The Causes of Molecular Evolution*. Oxford University Press, New York.
- [20] Gillespie, J.H. (1998) *Population Genetics: A Concise Guide*. The Johns Hopkins University Press, Baltimore, MD.
- [21] Graur, D. and Li, W.-H. (2000) *Fundamentals of Molecular Evolution*. Sinauer, Sunderland, MA.
- [22] Haldane, J.B.S. (1924) A mathematical theory of natural and artificial selection, Part I. *Trans. Camb. Phil. Soc.* **23**, 19-41.
- [23] Haldane, J.B.S. and Jayakar, S.D. (1963) Polymorphism due to selection of varying direc- tion. *J. Genet.* **58**, 237-242.
- [24] Hardy, G.H. (1908) Mendelian proportions in a mixed population. *Science*, **28**, 49-50.
- [25] Hopfield, J.J. (1978) Origin of the genetic code: A testable hypothesis based on tRNA structure, sequence, and kinetic proofreading. *Proc. Natl. Acad. Sci. USA*, **75**, 4334-4338.

- [26] Hopfield, J.J. (1994) Physics, computation, and why biology looks do different. *J. Theoret. Biol.* **171**, 53-60.
- [27] Hoppensteadt, F.C. and Peskin, C.S. (1992) *Mathematics in Medicine and the Life Sciences*. Springer-Verlag, New York.
- [28] Karlin, S. and Feldman, M.W. (1969) Linkage and selection: new equilibrium properties of the two-locus symmetric viability model. *Proc. Natl. Acad. Sci. USA*, **62**, 70-74.
- [29] Kimura, M. (1955) Solution of a process of random genetic drift with a continuous model. *Proc. Natl. Acad. Sci. USA*, **41**, 144-150.
- [30] Kimura, M. (1956) A model of a genetic system which leads to closer linkage under natural selection. *Evol.* **10**, 278-287.
- [31] Kimura, M. (1983) *The Neutral Theory of Molecylar Evolution*. Cambiridge Univ. Press, London.
- [32] Kimura, M., Kallianpur, G., and Hida, T. (1987) *Stochastic Methods in Biology*, Lecture Notes in Biomathematics, Vol. 70, Spinger-Verlag, New York.
- [33] Kolmogorov, A. (1935) Deviations from Hardy's formula in partial isolation. *C. R. Acad. Sci. USSR*, **3**, 129-132.
- [34] Kot, M. (2001) *Elements of mathematical ecology*. Cambridge Univ. Press, New York.
- [35] Kushner, H. (1967) *Stochastic Stability and Control*. Academic Press, New York.
- [36] Lewontin, R.C. and Kojima, K. (1960) The evolutionary dynamics of complex polymorphisms. *Evol.* **14**, 458-472.
- [37] Malécot, G. (1970) *The Mathematics of Heredity*, W.H. Freeman, San Francisco.
- [38] Mandl, P. (1968) *Analytical Treatment of One-Dimensional Markov Processes*. Springer-Verlag, New York.

- [39] Maruyama, T. (1983) Stochastic theory of population genetics. *Bullet. Math. Biol.* **45**, 521-554.
- [40] Nei, M. (1987) *Molecular Evolutionary Genetics*. Columbia Univ. Press, New York.
- [41] Nei, M. (1996) Phylogenetic analysis in molecular evolutionary genetics. *Ann. Rev. Genet.*, **30**, 371-403.
- [42] Nei, M. and Graur, D. (1984) Extent of protein polymorphism and the neutral mutation theory. *Evol. Biol.*, **17**, 73-118.
- [43] Nei, M., Maruyama, T., and Chakraborty, R. (1975) The Bottleneck effect and genetic variability in populations. *Evol.* **29**, 1-10.
- [44] Page, K.M. and Nowak, M.A. (2002) Unifying evolutionary dynamics. *J. Theoret. Biol.* **219**, 93-98.
- [45] Qian, H., Qian, M., and Tang, X. (2002) Thermodynamics of the general diffusion process: time-reversibility and entropy production. *J. Stat. Phys.* **107**, 1129-1141.
- [46] Shahshahani, S. (1979) A new mathematical framework for the study of linkage and selection. *Memoirs Am. Math. Soc.*, **17**, 1-34.
- [47] Tan, W.Y. (2002) *Stochastic Models with Applications to Genetics, Cancers, AIDS and Other Biomedical Systems*, World Scientific, Singapore.
- [48] Tavaré, S. (1979) A note on finite homogeneous continuous-time Markov chains. *Biometrics*, **35** 831-834.
- [49] Tavaré, S. (1980) Time reversal and age distributions, I. Discrete-time Markov Chains. *J. Appl. Prob.* **17**, 33-46.
- [50] van Nimwegen, E., Crutchfield, J.P., and Huynen, M. (1999) Neutral evolution of mutational robustness. *Proc. Natl. Acad. Sci. USA*, **96**, 9716-9720.

[51] Wright, S. (1937) The distribution of gene frequencies in populations. *Proc. Natl. Acad. Sci. USA*, **23**, 307-319.

[52] Wright, S. (1967) "Surfaces" of selective value. *Proc. Natl. Acad. Sci. USA*, **58**, 165-172.