Contents

15 Combinations of Forces

15.1 Mutation & Selection

Mutation - Review

If A mutates to B at rate u and B back to A at rate v, then

$$p_A = \frac{v}{u+v}$$

at equilibrium.

Note. A and B are two alleles at a single locus.

Mutation and Selection

Because there are more ways to make a bad protein than a good, functional protein, it would seem that mutation generally pushes toward a worse equilibrium. How does nature handle/control mutation?

DNA Repair: `http://www.ncbi.nlm.nih.gov/books/bv.fcgi?rid=hmg.section.1151`

15.1.1 Haploid

Haploid Model

Consider two alleles (genotypes) A and B. Let genotype A have allele frequency p at birth. We will consider how selection followed by mutation, in the circle of life, will impact the allele frequency.

If the fitness of genotype A is 1 and the fitness of genotype B is $1 - s$, then after selection the genotype frequency will shift from the initial p to

$$p^* = \frac{p}{p + (1-s)(1-p)} = \frac{p}{1-(1-p)s}$$

after selection.

Next, the individuals copy themselves and the possibility of mutation is introduced. The allele frequency after mutation is

$$p' = (1-u)p^*,$$

where we have neglected back mutation.

Haploid Equilibrium

Combining selection and mutation, we have

$$p' = \frac{(1-u)p}{1-(1-p)s}.$$

At equilibrium $p' = p$. Make the substitution and solve for the equilibrium,

$$\begin{aligned} p\left[1-(1-p)s\right] &= p(1-u) \\ up - sp(1-p) &= 0 \\ p\left[u - s(1-p)\right] &= 0 \end{aligned}$$

When will the system be at equilibrium?

$$q_e = \frac{u}{s}$$

where $q_e = 1 - p_e$.

If $s \gg u$, then q_e is predicted to be small. There won't be much allele B around despite the efforts of mutation to increase its numbers!

The force of selection is generally much stronger than the force of mutation.

Only at the DNA level can you see mutations that may have little impact on fitness and hence have small s, even on the order of magnitude of u.

What happens when $u > s$?

Example

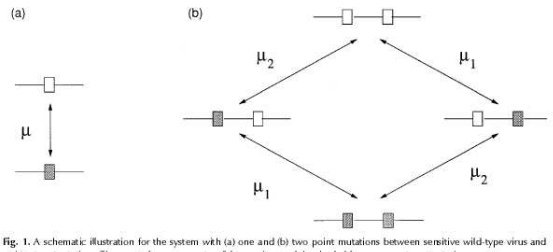


Fig. 1. A schematic illustration for the system with (a) one and (b) two point mutations between sensitive wild-type virus and resistant mutant virus. The empty box represents wild-type virus and the shaded box represents mutant virus.

Ribeiro et al. 1998. The frequency of resistant mutant virus before antiviral therapy. AIDS. **12**:461–465.

$$\begin{aligned}
\dot{x} &= \lambda - dx - \beta x\left[w + (1-s)m\right] \\
\dot{w} &= \beta x\left[(1-\mu)w + (1-s)\mu m\right] - aw \\
\dot{m} &= \beta x\left[\mu w + (1-s)(1-\mu)m\right] - am
\end{aligned}$$

$$\begin{aligned}
x_{t+1} - x_t &= \lambda - dx_t - \beta x_t\left[w_t + (1-s)m_t\right] \\
w_{t+1} - w_t &= \beta x_t\left[(1-\mu)w_t + (1-s)\mu m_t\right] - aw_t \\
m_{t+1} - m_t &= \beta x_t\left[\mu w_t + (1-s)(1-\mu)m_t\right] - am_t
\end{aligned}$$

Parameter	Meaning
x_t	Count of susceptible cells at generation t
w_t	Count of normal-infected cells in generation t
m_t	Count of mutant-infected cells in generation t
λ	Count of new susceptible cells born in each generation
d	Prob. susceptible cell dies in a generation
β	Prob. encounter susc. cell & normal-infected cell infects susc. cell
$(1-s)\beta$	Prob. encounter susc. & mutant-infected cell infects susc. cell
μ	Mutation rate
a	Prob. of death of infected cell

They find at equilibrium $\frac{m_e}{w_e} = \frac{\mu}{s}$. Does this look familiar?

15.1.2 Diploid

Recessive Model

First consider recessive mutants such that AA and AB individuals have relative fitness $w_A = 1$, while the homozygous recessive mutant BB has decreased relative fitness $w_B = 1 - s$.

Assume that the frequency of allele A is p in the starting gamete pool.

After selection

$$\begin{aligned}
p^* &= \frac{p^2 + p(1-p)}{p^2 + 2p(1-p) + (1-s)(1-p)^2} \\
&= \frac{p}{1 - s(1-p)^2}.
\end{aligned}$$

After mutation, where again we neglect back mutation

$$p' = \frac{p(1-u)}{1 - s(1-p)^2}$$

Recessive - Equilibrium

To find the mutation/selection balance for the recessive diploid case, again assume equilibrium so $p' = p = p_e$. Then,

$$1 - s(1-p_e)^2 = 1 - u$$

Therefore, at equilibrium

$$q_e = 1 - p_e = \sqrt{\frac{u}{s}}.$$

Comparing to the haploid case indicates that the gene frequency of the mutant allele B will be higher in the diploid recessive case than the haploid case. Can you explain this biologically?

Exercise. What is the frequency of affected individuals at equilibrium?

Example

Cystic fibrosis is a disease caused by a recessive allele, we shall call B. The frequency of affecteds at at birth is 1 in 2,500.

What is the cystic fibrosis allele B frequency q in the population?

Until recently, cystic fibrosis was fatal before affected individuals reached reproductive age, therefore $v_{BB} = 1 - s = 0$, where $s = 1$. At equilibrium

$$q_e = \sqrt{u}.$$

Is the required mutation rate reasonable for a disease caused by a single mutant protein?

Multiplicative Dominant

Now we consider the case where the mutant allele is completely or partially dominant to the wild type allele.

First, we assume geometric (multiplicative fitness).

AA	AB	BB
1	$1-s$	$(1-s)^2$

After selection

$$\begin{aligned} p^* &= \frac{\bar{w}_A}{\bar{w}} \\ &= \frac{p}{1-(1-p)s}, \end{aligned}$$

the same as the haploid case. Mutation affects allele frequencies in diploids just the same way it does in haploids, so the haploid results apply to diploid loci under multiplicative selection and $q_e \approx \frac{u}{s}$.

Selection in Homozygotes vs. Heterozygotes

Every generation a fraction $2sq(1-q)$ of heterozygotes are "killed" by selection. Each killing destroys a mutant B allele.

Equivalently, a fraction $q^2\left[1-(1-s)^2\right]$ homozygotes are "killed" by selection each generation. All these killings destroy 2 B alleles. The ratio of mutant alleles destroyed in heterozygotes to homozygotes is

$$\frac{2sq(1-q)}{2q^2\left[1-(1-s)^2\right]} = \frac{s(1-q)}{q(2s-s^2)} = \frac{1-q}{q(2-s)},$$

which falls between $\frac{1-q}{2q}$ and $\frac{1-q}{q}$ because $s \in [0,1]$.

Since $q \ll 1$, we conclude that this ratio is very large, and most mutants alleles are destroyed by selection on heterozygotes. That's because most mutant alleles are present in heterozygotes when the mutant allele is rare.

Partially Dominant

Let's parameterize partial dominance as follows:

AA	AB	BB
1	$1-hs$	$1-s$

h indicates how much of the fitness detriment in homozygote mutants BB is also shared by the heterozygote mutant carriers AB. So, for example, if $h \approx 1$, then heterozygotes are nearly as affected as homozygotes.

Because the mutant allele B will be rare when $s \gg u$, homozygote BB will be rare in the population. Selection will mostly be acting on heterozygotes, so there cannot possibly be much practical difference between the above selection scheme and

AA	AB	BB
1	$1-hs$	$(1-hs)^2$

The latter selection scheme is the familiar multiplicative.

Therefore, the mutant allele frequency for the general partial dominance fitness landscape is

$$q_e \approx \frac{u}{hs}.$$

Caution. The above result is true only when the mutant allele B is rare, i.e. $u \ll hs$.

Even fairly moderate heterozygote effects, e.g. h small, can still maintain mutant allele frequency q_e low as long as $u \ll s$.

Counterintuitively, as far as population impact, the small fitness effects on mutant carriers (heterozygote AB) is much more important than the potentially huge impacts on affected homozygotes BB.

15.1.3 Theoretical Predictions

Haldane-Muller Principle - Haploids

We will now concern ourselves with the magnitude of the detrimental effect of mutation on a population.

For haploids, the mean relative fitness of the population is

$$\bar{w} = 1 - q + (1-s)q = 1 - sq$$

and at equilibrium $q = q_e = u/s$, so mean relative fitness is

$$\bar{w}_e = 1 - u.$$

Surprisingly, the effect of mutation on the mean relative fitness of the population is to decrease it by fraction u, which is **independent of the fitness of the mutant!**

Haldane-Muller Principle - Diploids

For diploids, we have mean relative fitness of a recessive allele is

$$\bar{w} = (1-q)^2 + 2q(1-q) + (1-s)q^2 = 1 - sq^2$$

At equilibrium $q_e = \sqrt{u/s}$, so again

$$\bar{w}_e = 1 - u.$$

For partial or fully dominant alleles, mean relative fitness is

$$\bar{w} = (1-q)^2 + (1-hs)q(1-q) + (1-s)q^2 = 1 - 2hsq(1-q) - sq^2.$$

At equilibrium $q_e \approx \frac{u}{hs}$,

$$\bar{w}_e = 1 - 2u + \frac{2u^2}{hs} - \frac{u^2}{h^2 s} \approx 1 - 2u,$$

since u is very small and u^2 is neglible.

Genetic Load

The fraction of mean relative fitness lost because of mutation is called *genetic load*. It represents the cost of mutation.

$$L = \frac{w_{\max} - \bar{w}}{w_{\max}},$$

where $w_{\max}$ is the fitness of the maximally fit genotype in the population. If $w_{\max} = 1$, then $L \approx 2u$ for diploids.

Consider n independent loci each mutating and contributing to the load. If we assume fitness effects across loci are multiplicative, then for n partially dominant loci, the mean fitness is

$$(1-2u)^n \approx e^{-2un}$$

and genetic load is $1 - e^{-2un}$.

Exercise. Show the cost of recessive mutations is less than the cost of dominant mutations.

15.2 Migration & Selection

Migration - Review

Whalund Effect

$$P_{AA}^{(1)}(t+1) = [\mathrm{E}(p)]^2 + \mathrm{Var}[p(t)].$$

In words, the frequency of homozygotes in a population receiving immigrants from multiple population is increased over Hardy-Weinberg expected frequencies by a factor equalling the variance in allele frequencies across different populations and weighted by their relative contributions.

As before, we are now interested in how migration interacts with other forces. In particular, selection. Migration, like mutation acts to homogenize populations. Selection is the only force that can act in the opposite direction. For spatial migration, selection can lead to more spatial heterogeneity when the most fit variant depends on geographic location.

15.2.1 One-Island Model

Haploid

Suppose in the one-island model allele B is fixed on the continent (allele A has been eliminated by selection against it). On the island, allele A is favored by selection. Let the relative fitnesses on the island be 1 for A and $1 - s$ for B.

If there were no immigration, allele A would dominate on the island and would achieve mutation/selection balance. Immigration disrupts equilibrium. How much immigration is allowed while still maintaining allele A on the island?

Let the allele frequency of A start at p on the island. After selection, the allele frequency of allele A is

$$p^* = \frac{p}{1-(1-p)s}.$$

After migration, the allele frequency becomes

$$p' = (1-m)p^* + m \times 0 = (1-m)p^* = \frac{p(1-m)}{1-(1-p)s}.$$

We recognize the equation as the one for mutation/selection balance in a haploid population. We know that at equilibrium either $p_e = 0$ or $p_e = 1 - \frac{m}{s}$, where migration rate m has replaced mutation rate u.

When $m \geq s$, then the second equilibrium predicts $p_e \leq 0$, so $p_e = 0$ is the only valid equilibrium. In other words, migration will overwhelm the efforts of selection and will eliminate allele A.

When $m < s$, $p_e = 1 - \frac{m}{s}$ will be a stable equilibrium (not shown). In this case, selection dominates migration and is able to maintain some beneficial A allele on the island.

Note, that the island population is experiencing fitness degradation because of immigration from the island. It would be to their benefit to not intermix with incoming immigrants and, assuming the immigrants compete for the same resources, even prevent them from arriving!

Diploid Dominant

Now consider a diploid population. We need to carefully define *when* immigration happens. Suppose selection occurs first, then migrants arrive.

We consider the diploid dominant where AA and AB individuals have fitness 1 and BB individuals have fitness $1-s$.

Then the update equations for the allele frequency of allele A are

$$p' = \frac{p(1-m)}{1-s(1-p)^2}.$$

and the change in one generation is

$$p' - p = \frac{p\left(-m + s(1-p)^2\right)}{1-s(1-p)^2}.$$

The change is positive if

$$m < s(1-p)^2.$$

$$m < s(1-p)^2$$

When $m > s$, the above condition cannot be met and allele A will decline in the population until it is eliminated.

When $m < s$, selection is able to control migration and there is a stable equilibrium

$$p_e = 1 - \sqrt{\frac{m}{s}}.$$

Diploid Recessive

Now consider a recessive allele such that AA individuals have relative fitness 1 and AB and BB individuals have relative fitness $1-s$. The update equations for allele frequency of A with migration after selection is

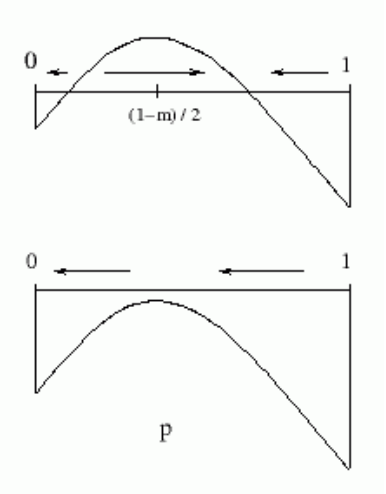
$$p' = \frac{p(1-m)\left[1-(1-p)s\right]}{1-\left[2p(1-p)+(1-p)^2\right]s},$$

leading to per-generation change of

$$\begin{aligned} p' - p &= \frac{p\left\{(1-m)\left[1-(1-p)s\right] - 1 + \left[2p(1-p)+(1-p)^2\right]s\right\}}{1-\left[2p(1-p)+(1-p)^2\right]s} \\ &= \frac{p\left\{(1-m)\left[1-(1-p)s\right] - \left[1-s(1-p^2)\right]\right\}}{1-\left[2p(1-p)+(1-p)^2\right]s}. \end{aligned}$$

Whether the change is negative, positive, or there is no change depends on

$$(1-m)\left[1-(1-p)s\right] <=> \left[1-s(1-p^2)\right].$$



- $p = 0$ is always a stable equilibrium.
- There are two additional equilibria when $s^2(1-m)^2 - 4ms(1-s) > 0$. When $s \ll 1$ and $m \ll 1$, then there are two equilibria if $s > 4m$. Only the larger of the two is a stable equilibrium.
- Equilibrium $p = 0$ will attract if allele frequency of A is too low.
- Equilibrium $p > 0$ will attract if allele frequency of A is high enough.

Locally Adaptive Recessive Alleles

- Locally advantageous alleles (alleles that are favored in limited geographic areas) are not expected to be recessive alleles.
- If there are locally favored recessive alleles, they should not appear at frequencies less than $\frac{1}{2}$. To see this, consider the maximum of the parabola ($p = \frac{1-m}{2} \approx \frac{1}{2}$).

Summary

- When migration dominates selection, a locally adaptive allele cannot be maintained in the local population.
- When selection dominants migration, a haploid or diploid dominant allele can be maintained in the local population.
- When selection dominants migration, a diploid recessive allele can only be persist in the local population if it is initially present at high frequency (above about $\frac{1}{2}$).

15.3 Mutation & Drift

Genetic Drift Compared to Other Forces

A fundamental question:

When do the deterministic forces of mutation, selection, and migration overcome the random variation caused by genetic drift.

- Is visible diversity a reflection of selection/mutation balance or random variation introduced by finite population sizes?
- Is evolution dominated by random events or a purposeful forward march dictated by selective forces?

15.3.1 Infinite Isoalleles Model

Infinite Isoalleles Model

Assume a Wright-Fisher model with N diploid individuals. Recall this model applies very closely, even to non-Wright populations, when

- The sex ratio does not depart substantially from 1.
- The population size N is not so small that selfing becomes frequent.

The *infinite isoalleles model* introduces mutation into the mix and assumes

- Each mutation produces a novel, never-before-seen allele.
- Each mutant allele is selectively neutral (no effect on viability and fitness).

We can proceed to study this system in one of two equivalent ways

- Follow the amount of inbreeding in the population over time.
- Compute the variance in allele frequency between multiple, replicate populations.

Exercise: How many alleles are one mutational step away from an allele at a locus of length 3000 nucleotides? How many alleles are two mutational steps away?

Equilibrium

$$f_{t+1} = \frac{1}{2N} + \left(1 - \frac{1}{2N}\right) f_t$$

If each new mutation produces an allele that is new, then recipients of this new mutant cannot be inbred, no matter where or how they get their second allele. An inbred individual is created in the usual way *and* if neither allele is mutated when transmitted to the next generation:

$$f_{t+1} = (1-u)^2 \left[\frac{1}{2N} + \left(1 - \frac{1}{2N}\right) f_t\right].$$

In the presence of mutation f_t no longer rises inexorably to 1. Instead it will achieve some equilibrium value between 0 and 1 as long as N is finite and u is not zero.

To get a simple expression, we make two approximations

- Assume u is very small and ignore u^2 terms: $(1-u)^2 \approx 1 - 2u$
- Also assume N is quite large and ignore $\frac{u}{N}$ terms.

Thus,

$$\begin{aligned} f_{t+1} &\approx (1-2u)\left[\frac{1}{2N} + \left(1 - \frac{1}{2N}\right) f_t\right] \\ &\approx \frac{1}{2N} + \left(1 - \frac{1}{2N}\right) f_t - 2u f_t. \end{aligned}$$

Rearranging to get a recurrence relation for heterozygosity, we have

$$h_{t+1} \approx \left(1 - \frac{1}{2N}\right) h_t + 2u\,(1 - h_t)$$

The change in a single generation is

$$\Delta h_t \approx -\frac{1}{2N} h_t + 2u\,(1 - h_t)$$

At equilibrium $\Delta h_t = 0$, yielding equilibrium heterozygosity

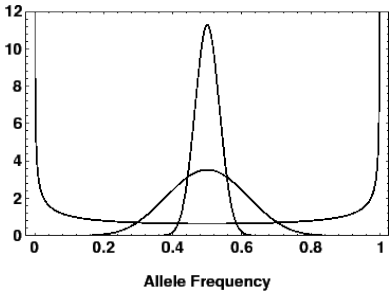
$$h \approx \frac{4Nu}{1 + 4Nu}$$

and

$$f = 1 - h \approx \frac{1}{1 + 4Nu}.$$

Conclusions

- Mutation affects inbreeding only through the product $4Nu$.
- The higher the mutation rate, the slower the accumulation of inbreeding.
- Doubling the mutation rate is as effective at reducing inbreeding as doubling the population size.



Consider multiple, isolated subpopulations all with the same breeding structure, size and forward *and* reverse mutation rate u. Which of these plots describes the distribution of an allele frequency across these populations when $4Nu$ is large? Small?

15.3.2 Finite Isoalleles Model

Finite Isoalleles Model

Clearly the infinite isoalleles model is unrealistic in assuming that all mutations produce new variants. What happens when there is only a finite number of possible mutants (e.g. consider a single nucleotide position that can assume 1 of 4 possible states, A, C, G or T)?

Now assume that there are only K possible alleles. Again, prior to mutation two alleles in generation t will be IBD with probability $\frac{1}{2N} + \left(1 - \frac{1}{2N}\right) f_{t-1}$. After mutation, there are three possibilities

- Neither mutates with probability $(1-u)^2$ and they are still IBD.
- Only one mutates with probability $u(1-u)$ and they are no longer IBD.
- Both mutate in this generation, then ??

We must define what happens to alleles that are originally different and then mutate to the same allele. We will say they become ibd alleles. This definition contrasts with our previous assumption, but allows us to better relate observable *state* with unobservable *IBD state*. So, when two IBD alleles both mutate

- They are still IBD if they mutate to the same allele with probability $\frac{u^2}{K-1}$.
- They become non-IBD with probability $u^2 \frac{K-2}{K-1} \frac{1}{K-1}$.

Alleles that are NOT ibd (with probability $\left(1 - \frac{1}{2N}\right)(1 - f_t)$) can arise by mutation:

- They both mutate to the same allele with probability $u^2 \frac{K-2}{K-1} \frac{1}{K-1}$.
- One mutates to match the first with probability $\frac{2u(1-u)}{K-1}$.

Taken together, the update equation becomes

$$\begin{aligned} f_{t+1} &= \left[(1-u)^2 + \frac{u^2}{K-1}\right]\left[\frac{1}{2N} + \left(1 - \frac{1}{2N}\right) f_t\right] \\ &\quad + \left[\frac{2u(1-u)}{K-1} + \frac{u^2(K-2)}{(K-1)^2}\right]\left[\left(1 - \frac{1}{2N}\right)(1 - f_t)\right]. \end{aligned}$$

Approximate Model

Immediately we want to start simplifying, and we do so by neglecting small terms. To start, drop all terms of order u^2 or u/N to yield

$$f_{t+1} \approx \frac{1}{2N} + \frac{2u}{K-1} + f_t\left[1 - \frac{1}{2N} - 2u\left(\frac{K}{K-1}\right)\right].$$

At equilibrium, of course, $f_{t+1} = f_t = f$, so making the substitution and solving produces equilibrium inbreeding of

$$f \approx \frac{1 + \frac{4Nu}{K-1}}{1 + \frac{4NuK}{K-1}},$$

a good approximation when u is small and N is not too small.

- As $K \to \infty$, $f \to \frac{1}{1+4Nu}$, the equation for the infinite isoalleles model.
- As $4Nu \to \infty$, $f \to \frac{1}{K}$, and alleles will only be IBD if they are the same allele.

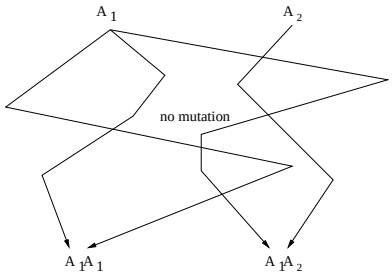
Infinite vs. Finite Isoalleles Model

$4Nu =$ \ $K =$	2	4	8	16	∞
0.1	0.9167	0.9118	0.9103	0.9096	0.9091
0.5	0.75	0.70	0.6818	0.6739	0.6667
1.0	0.667	0.5714	0.5333	0.5161	0.5000
10.0	0.5238	0.3023	0.1954	0.1429	0.0909

There is little effect of the finite allele model as long as K is not too small and $4Nu$ is not too big.

This is not to say there aren't problems with the infinite isoalleles model, e.g. DNA evolution.

Homozygosity and Infinite Isoalleles Model



In addition, it is important to note that as time passes, f_t in the infinite isoalleles model is a measure of homozygosity. An individual will be homozygous if s/he received two copies of the same allele or if s/he received to copies of different copies of the same allele from the base population. However, as time passes most every allele present in the base population will have undergone mutation, so the second possibility becomes increasingly unlikely.

Rate of Mutation

While we can now predict the level of inbreeding (homozygosity in the infinite alleles model) there is of course ongoing mutation, a continual flux of alleles from one allele to the next novel allele to the next...ad nauseum. Some alleles will fix (spread) into the population. Others will be eliminated. How rapidly do new alleles appear and fix in the population? What is the population level flux of new mutations?

On average $2Nu$ new mutants are introduced in each generation. Each alleles will become fixed (in the absence of mutation) with probability $\frac{1}{2N}$ since there is no selection and an allele's probability of fixation only depends on its initial frequency in the population.

In the presence of mutation, the fixing allele may mutate before it completely fixes, but in any case, a whole new mutant or set of mutants will sweep through when its descendents finally fix.

Overall then, the population level flux of alleles is

$$\frac{2Nu}{2N} = u.$$

15.3.3 Bottlenecks

Bottlenecks, Inbreeding, and Mutation

Previously we indicated that population bottlenecks introduced permanent inbreeding into populations that could not be eliminated even if the population were to expand to infinite size.

Mutation *can* reverse inbreeding by reintroducing variability into the population after a bottleneck eliminates it.

Imagine a large population with very little inbreeding (f_t is small). Now shrink that population to small size N. Then in the next generation after shrinking the population, the inbreeding will be

$$\begin{aligned} f_{t+1} &= (1-u)^2 \left[\frac{1}{2N} + \left(1 - \frac{1}{2N}\right) f_t \right] \\ &\approx \frac{(1-u)^2}{2N} \approx \frac{1}{2N}, \end{aligned}$$

when u is small.

A new equilibrium for the new population size N should be established fairly quickly. Assume the new equilibrium is reached at

$$f \approx \frac{1}{1+4Nu},$$

which for small u and N is very close to 1.

Now, suppose the population expands back to a large population size. In the first generation, the new inbreeding level will be

$$\begin{aligned} f_{t+1} &= (1-u)^2 \left[\frac{1}{2N} + \left(1 - \frac{1}{2N}\right) f_t \right] \\ &\approx (1-u)^2 \approx 1 - 2u. \end{aligned}$$

We conclude that each generation f_t is only decreased by the small amount $2u$ per generation early after the bottleneck is released.

Real Examples of Population Bottlenecks

Thus, bottlenecks leave a lasting impression on populations, even those subject to mutation. One need not look far to find examples of famous bottlenecks

- The global human population appears to have experienced approximately 70000 years ago, perhaps when the Toba supervolcano erupted in Indonesia and changed the world's climate.
- American bison were nearly extinct in 1890, but are making a comeback in places like Yellowstone.
- Northern elephant seal populations dropped to 30 in the 1890's, but if you go to the California coastline in winter, be sure and visit with the growing numbers.

The first example is suggested by DNA evidence alone since no census numbers for humans 70000 years ago are available.

Accomodating More Non-Wright-Fisher Models

Many types of non-Wright-Fisher models can be modeled using the same equations by using the effective population size N_e instead of the census population size (where we gave some examples of how to compute N_e for some special cases).

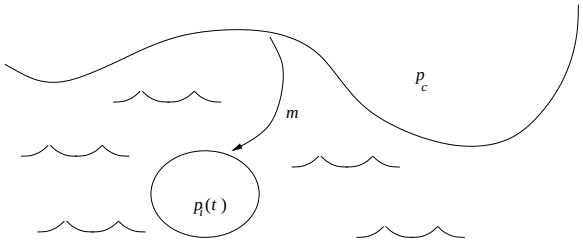
Thus, the critical role of $4Nu$ is replaced by $4N_eu$ and all is well.

Among others who have worked on these generalizations to more complex models is Iowa State's own Dr. Pollak.

15.4 Migration & Drift

15.4.1 One-Island Model

One-Island Model - Mutation & Drift



If we assume migrants arrive as gametes and just contribute to the infinite pool, then if p_t is the frequency of allele A among the gametes before the immigrants arrive, then

$$p_t^* = (1-m)p_t + mp_c$$

is the frequency of allele A after the immigrant gametes contribute to the pool, where p_c is the frequency of the A allele on the continent.

Mean Allele Frequency

The adult allele frequency results after random union of gametes and random kill-off of excess individuals. The expected allele frequency in the next generation is

$$\mathrm{E}(p_{t+1}) = \mathrm{E}(p_t^*) = (1-m)\mathrm{E}(p_t) + mp_c,$$

where everything on the right side is constant except p_t which fluctuates around because of random drift.

At stationarity, there is no longer change in the expected allele frequency of on the island $\mathrm{E}(p_{t+1}) = \mathrm{E}(p_t)$ (though of course random fluctuations will continue). The one-island model may not start at equilibrium (for example imagine an island with no A alleles with immigration from a continent with only A alleles).

$$\mathrm{E}(p) = (1-m)\mathrm{E}(p) + mp_c,$$

so $\mathrm{E}(p) = p_c$.

Variance in Allele Frequency

We now investigate the variability of p_t around its equilibrium expectation p_c. This is the variability that would be observed across multiple islands all connected by migration to the same continent (but not each other) or a measure of variability around the mean in a time series of the island allele frequency over time.

Consider $x_t = p_t - p_c$ the deviation from equilibrium at generation t. Notice and remember that $E[x_t] = 0$ from the previous slide.

Since p_c is constant $\mathrm{Var}(x_t) = \mathrm{Var}(p_t)$. Then, from

$$p_t^* = (1-m)p_t + mp_c$$

and $p_c + x_t = p_t$, we have

$$p_c + x_t^* = (1-m)(p_c + x_t) + mp_c$$

Now x_t^* is the deviation in proportion of A alleles after migration, but before drift, so

$$x_t^* = (1-m)x_t$$

where x_t is the deviation in proportion of A alleles before migration and genetic drift.

But, x_{t+1} comes from x_t^* by random sampling from an infinitely large pool of gametes, where the proportion of gametes in the pool is x_t^*. Therefore, we can write

$$x_{t+1} = x_t^* + e_t.$$

where e_t is the randomness introduced by this selection process. Notice as before, that e_t has expectation 0 and is not correlated with x_t^*. In the end then we have the following update equation

$$x_{t+1}^* = (1-m)(x_t^* + e_t).$$

On our way to the variance we stop by $\mathrm{E}(x_{t+1}^{*2})$ (as usual), but

$$\begin{aligned}
\mathrm{E}(x_{t+1}^{*2}) &= (1-m)^2 \mathrm{E}\left[(x_t^* + e_t)^2\right] \\
&= (1-m)^2 \mathrm{E}\left[x_t^{*2} + 2e_t x_t^* + e_t^2\right] \\
&= (1-m)^2 \left[\mathrm{E}(x_t^{*2}) + 2\mathrm{E}(e_t)\mathrm{E}(x_t^*) + \mathrm{E}(e_t^2)\right] \\
&= (1-m)^2 \left[\mathrm{E}(x_t^{*2}) + \mathrm{E}(e_t^2)\right].
\end{aligned}$$

Now, examine the second term

$$\begin{aligned}
\mathrm{E}(e_t^2) &= \mathrm{E}\left[\mathrm{E}(e_t^2 \mid p_t^*)\right] \\
&= \mathrm{E}\left[\frac{p_t^*(1-p_t^*)}{2N}\right] \\
&= \mathrm{E}\left[\frac{(p_c + x_t^*)(1-p_c-x_t^*)}{2N}\right].
\end{aligned}$$

Continuing, we have

$$\begin{aligned}
\mathrm{E}(e_t^2) &= \mathrm{E}\left[\frac{(p_c + x_t^*)(1-p_c-x_t^*)}{2N}\right] \\
&= \frac{1}{2N}\mathrm{E}\left[p_c(1-p_c) + x_t^*(1-p_c) - p_c x_t^* - x_t^{*2}\right] \\
&= \frac{p_c(1-p_c)}{2N} - \frac{1}{2N}\mathrm{E}(x_t^{*2}).
\end{aligned}$$

And if we write V_t for $\mathrm{E}(x_t^{*2})$ and put everything back together, we have

$$V_{t+1} = (1-m)^2 \left[V_t + \frac{p_c(1-p_c)}{2N} - \frac{V_t}{2N}\right].$$

Now, we seek the equilibrium value $\lim_{t\to\infty} V_t = V$.

$$V = \frac{p_c(1-p_c)}{2N - (1-m)^2(2N-1)}.$$

Approximation

When m^2 and $\frac{m}{N}$ are small, we have

$$V \approx \frac{p_c(1-p_c)}{4Nm+1}.$$

- When $m = 0$, there is no migration and $V = p_c(1-p_c)$, consistent with previous results obtained when all populations fix either on A or non-A.
- As $4Nm$ increases, the island frequency will show little variation around the continental frequency p_c.
- Note that we expect $2Nm$ alleles in the adult population to have come from immigrants and therefore roughly Nm individuals will migrate each generation. This is how you relate the results of this model to a model where adults immigrate (though the derivation doesn't quite work for adults).

General Rules of Thumb

Mutation

If $2Nu \gg 1$, substantial genetic variability will be maintained in a population, counteracting the effects of inbreeding.

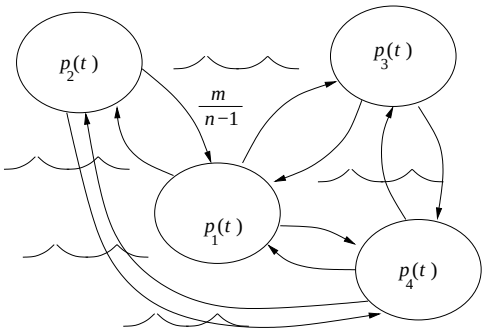
Migration

One migrant per generation is generally sufficient to stop the effects of genetic drift in a local population.

15.5 Migration, Mutation, & Drift

15.5.1 The Island Model

The Island Model



The model we will use here includes both migration and mutation. We assume adults produce infinite gametes, then migration replaces a proportion m of those gametes randomly from the other $n-1$ islands, then mutation alters the gametes, then random union creates adults that are severely parred down to just N surviving adults in the next generation.

The island model is the simplest model where we can ask:

- What level of migration can occur without homogenizing subpopulations?
- When do subpopulations start behaving as a single, panmictic population?

The forces at work in this system and their effects are:

- Drift drives isolated subpopulations apart since it is random and no two are likely to "drift" in the same direction. Drift is *removing* genetic variation from each island.
- Mutation under the infinite isoalleles model drives isolated subpopulations apart since the chance that two experience the same mutations is 0. In addition, mutation is diversifying the populations within each island.
- Migration is bringing the subpopulations closer together since it is the only force that increases the number of shared alleles. It is also diversifying each subpopulation by introducing novel mutants from other populations.

The Island Model Recurrence Relations

Let F_B (between) be the probability that two alleles drawn from different populations are identical. Let F_W (within) be the probability that two alleles drawn from the same population are identical.

Then, the recurrence relations for F_B and F_W are:

$$\begin{aligned}
F_W(t+1) &= (1-u)^2 \left\{ \left[(1-m)^2 + \frac{m^2}{n-1} \right] \left[\frac{1}{2N} + \left(1 - \frac{1}{2N} \right) F_W(t) \right] \right. \\
&\quad \left. + \left[1 - (1-m)^2 - \frac{m^2}{n-1} \right] F_B(t) \right\} \\
F_B(t+1) &= (1-u)^2 \left\{ \left[(1-m)^2 + 2m(1-m) \left(\frac{n-2}{n-1} \right) + m^2 \left(1 - \frac{n-2}{(n-1)^2} \right) \right] F_B(t) \right. \\
&\quad \left. + \left[\frac{2m(1-m)}{n-1} + m^2 \left(\frac{n-2}{(n-1)^2} \right) \right] \left[\frac{1}{2N} + \left(1 - \frac{1}{2N} \right) F_W(t) \right] \right\}
\end{aligned}$$

Remember: (gamete production) → (migration) → (mutation) → (random union) → (random survival of N).

Breakdown - Within Populations

Term	Probability
$(1-u)^2$	No mutation of two randomly selected alleles
$(1-m)^2$	Two randomly selected alleles are not new immigrants
$\frac{m^2}{(n-1)}$	Two alleles selected from one population are both new immigrants from the *same* population
$\frac{1}{2N} + \left(1 - \frac{1}{2N}\right) F_W$	New inbreeding within local (non-migrant) population given inbreeding a generation ago

Breakdown - Between Populations

Term	Probability
$\frac{1}{n-1} \times \frac{n-2}{n-1} = \frac{n-2}{(n-1)^2}$	2 randomly selected alleles are from the different populations, excluding 1 population
$\frac{m^2(n-2)}{(n-1)^2}$	2 new migrant alleles selected from 2 populations were in 1 population a generation ago
$\frac{2m(1-m)}{n-1}$	2 alleles from 2 populations, 1 new migrant the other not, were in 1 population a generation ago
$2m(1-m)\left(\frac{n-2}{n-1}\right)$	2 alleles from 2 populations, 1 new migrant the other not, were in 2 populations a generation ago
$m^2\left(1 - \frac{n-2}{(n-1)^2}\right)$	2 new migrant alleles from 2 populations were in 2 populations a generation ago

The Island Model Equilibrium

To find an equilibrium we can solve the two equations above by assuming $F_W(t+1) = F_W(t) = F_W$ and $F_B(t+1) = F_B(t) = F_B$, but it is a lot easier to first make some approximations. Neglect all terms involving $m^2, u^2, mu, m/N$, and u/N as being negligible.

$$\begin{aligned}
F_W &= \frac{1}{2N} + (1 - 2u - 2m - \frac{1}{2N})F_W + 2mF_B \\
F_B &= \left(1 - 2u - \frac{2m}{n-1}\right) F_B + \frac{2m}{n-1} F_W.
\end{aligned}$$

Use the second equation to compute

$$\frac{F_B}{F_W} = \frac{\frac{m}{n-1}}{u + \frac{m}{n-1}},$$

that tells us how similar genes are between populations with respect to how similar they are within populations.

The Island Model Equilibrium

$$\rho := \frac{F_B}{F_W} = \frac{\frac{m}{n-1}}{u + \frac{m}{n-1}},$$

Note:

- Completely unrelated populations with distinct allele contents have $F_B = 0$ and hence $\rho = 0$
- When populations are completely mixed (i.e. no subpopulations exist), $F_B = F_W$ and $\rho = 1$.

Letting ρ be defined as above and substituting the result of the second equation into the first, we obtain

$$F_W = \frac{1}{1 + 4Nu + 4Nm(1-\rho)},$$

Interpretation

$$F_W = \frac{1}{1 + 4Nu + 4Nm(1-\rho)},$$

The amount of inbreeding within subpopulations is decreased by a higher mutation rate and higher migration rate (regardless of the value of ρ). In other words, both mutation and migration (unless $\rho = 1$) diversify populations.

It is as if the mutation rate "increases" in the presence of migration by additive factor $m(1-\rho)$. Similarly the migration rate is supplemented by $u - m\rho$ in the presence of mutation.

ρ measures how homogeneous different populations appear with respect to how homogeneous they are internally. The larger ρ (up to maximum 1), the greater the similarity between subpopulations. When $\rho = 1$, migration can no longer homogenize (and also diversify the allele content of) subpopulations.

Increasing the island population size N (weakening genetic drift) allows the subpopulations to retain more variability and F_W is small. However, since ρ is constant with respect to N, we also know F_B is decreasing and the similarity between populations is not high either.

When the number of islands n increases $\rho \to 0$ indicating the islands are less and less similar, largely because there is hardly any migrant traffic between any two particular pairs of islands. The approximate formula becomes

$$F_W = \frac{1}{1 + 4N(u+m)},$$

and migration acts like another source of mutation introducing new mutants produced elsewhere but hardly homogenizing the populations at all.

The mutation rate u helps to maintain local diversity and works to differentiate populations (decreasing ρ).

Increasing migration homogenizes the populations (increasing ρ).

m reaches a maximum of $\frac{n-1}{n}$ when each individual is equally likely to come from any 1 of the n subpopulations (i.e. a randomly mating panmictic population of size nN. When this is true

$$F_B = F_W = \frac{1}{1 + 4Nnu}.$$

Absence of Mutation

What happens when there is no mutation in the island model?

Without mutation to restore variability inbreeding will gradually increase F_W, and migration will homogenize, therefore increasing F_B. In fact, when $u = 0$

$$\begin{aligned} \rho &= \frac{\frac{m}{n-1}}{u + \frac{m}{n-1}} = 1 \\ F_W &= \frac{1}{1 + 4Nu + 4Nm(1-\rho)} = \frac{1}{1+0+0} = 1 \end{aligned}$$

shows $F_W = F_B = 1$ at equilibrium. The whole population will eventually fix a single allele at each locus, but

- What is the rate of this fixation?
- Will the subpopulations be homogeneous during this fixation or will their be variability between subpopulations?

Rewrite the equations for F_W and F_B in terms of $H_W = 1 - F_W$ and $H_B = 1 - F_B$ to obtain

$$\begin{aligned} H'_W &= \left(1 - 2m - \frac{1}{2N}\right) H_W + 2mH_B \\ H'_B &= \frac{2m}{n-1} H_W + \left(1 - \frac{2m}{n-1}\right) H_B. \end{aligned}$$

Written in matrix notation, we have $H_{t+1} = MH_t$, where the transpose of H_t is $(H_W(t), H_B(t))$. The solution is $H_t = M^t H_0$. If M has eigenvalues λ_1 and λ_2 and eigenvectors y_1 and y_2, then $H_0 = c_1 y_1 + c_2 y_2$ for some constants c_1 and c_2. As a result

$$H_t = c_1 \lambda_1^t y_1 + c_2 \lambda_2^t y_2.$$

The eigenvalues $\lambda_1, \lambda_2 \in [0, 1]$ (This is not proven, but you know $H_t \in [0, 1]$, so it is a necessity in light of the above equation).

With $\lambda_1, \lambda_2 \in [0, 1]$ and knowing $\lambda_1 \neq \lambda_2$ (also not proven but true when the determinant of M is nonzero, which is the case when m and N are reasonable), it is not hard to see that after the first few iterations (after t is reasonably large) the value of H_t will be dominated by the term involving the larger eigenvalue. To be specific, suppose $\lambda_1 > \lambda_2$, then for reasonably large t, we know $\lambda_1^t \gg \lambda_2^t$ and we can neglect the second term.

After this time, H_t is increasing at geometric rate λ_1 since

$$H_t \approx c_1 \lambda_1^t y_1.$$

To determine the properties of this island system without mutation, we therefore need to know the largest eigenvalue of M. You can use computer software to find the eigenvalues for any M. Here, we consider some special cases.

Special Cases of Mutation Absence

- When Nm is large because the island population sizes N are very large or migration rate m is high, the population behaves as if it were a single, panmictic population and

  $$\lambda_1 \approx 1 - \frac{1}{2Nn}.$$

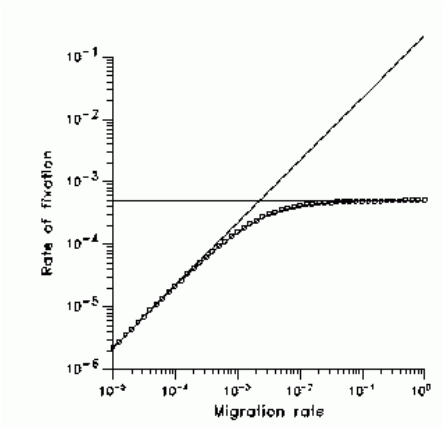
  The increase in inbreeding each generation is $\frac{1}{2Nn}$ and is independent of the migration rate m.

- When m is very small the rate of loss becomes

$$\lambda_1 \approx 1 - \frac{2m}{n-1}.$$

The increase in inbreeding each generation $\frac{2m}{n-1}$ is independent of the island population size N. It does depend on the migration rate m and the number of islands n. Or, viewed differently, it depends on $\frac{m}{n-1}$ the probability that a migrant immigrates from one other particular island.

The transition in the rate of loss is not gradual with changes in m. Plotting the exact value of $1 - \lambda_1$ for various levels of migration when $n = 10$ and $N = 100$, reveals a threshold migration rate. (Rate of fixation = Rate of new inbreeding):



The two lines are the approximations obtained at the extremes of low and high m from two slides ago. The transition occurs roughly where the two lines intersect, i.e. when

$$\frac{1}{2Nn} = \frac{2m}{n-1}$$

or

$$4Nm = \frac{n-1}{n}.$$

So, roughly, we obtain the principle that the population acts as if *not* subdivided when $4Nm > 1$ and otherwise acts as if completely subdivided (and not mixing) when $4Nm < 1$. The rapidity of the transition indicates that only values quite close to 1 are not easily categorized into these two cases.

A good approximation of the maximum eigenvalue is obtained as

$$\lambda_1 \approx 1 - \frac{1}{2Nn + \frac{n-1}{2m}}.$$

Degree of Differentiation Without Mutation

The only remaining question is whether the subpopulations show much differentiation during the time before fixation.

- When Nm is large, there is little geographic differentiation (it is behaving as a single population after all).
- When Nm is small, there is substantial geographic differentiation. The fixation of an allele in the whole population is much slower than the fixation of alleles in individual populations (which is never quite accomplished since migration will reintroduce lost alleles at a slow, but ever-present rate). In fact, waiting for fixation at the whole population is like waiting for the eventuality that all subpopulations fix on the same allele and while you know all subpopulations fix, they are not very likely to all fix on the same allele.

15.6 Selection & Drift

15.6.1 Introduction

Drift vs. Selection

The combination of selection and drift is difficult to model theoretically. The required mathematics will soon venture outside the depth of knowledge you are required to have for this course, so we will not be able to show all derivations. You will be responsible for knowing the main findings and implications.

The first lesson to learn is that natural selection introduces a bias in the change of allele frequencies over time. Selection *for* an allele will tend to cause its frequency p_t to increase in the next generation. However, genetic drift is still a random and unbiased force that can both help and hinder selection. Because of genetic drift, the frequency p_t may

- increase even more than expected, or
- actually decrease

in a generation.

We handled selection previously by taking the mean offspring per adult W_{AA}, W_{Aa}, and W_{aa} and defining the mean relative number of offspring of each genotype, e.g. $w_{AA} = \frac{W_{AA}}{W_{Aa}} = 1 + s$ implies an AA individual with will have on average $1 + s$ times as many offspring as an Aa individual.

In large populations this description was sufficient. We considered "average" individuals and that was enough.

Now, each individual of genotype g may have random variable $Y_g = 0, 1, 2, \ldots$ offspring with

$$\frac{\mathrm{E}(Y_{AA})}{\mathrm{E}(Y_{Aa})} = \frac{W_{AA}}{W_{Aa}} = w_{AA}$$

but the particular value of Y_g can have a tremendous impact on whether a new allele, selected or not, will survive and spread into a population or not.

15.6.2 Extinction Probability

Will a New Mutant Fix?

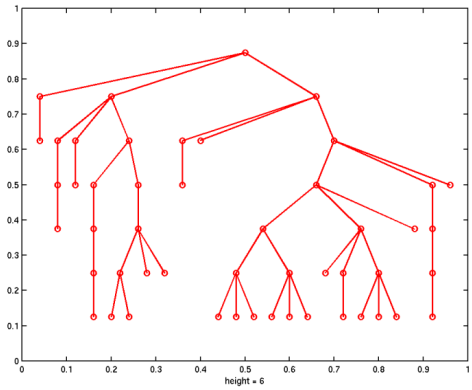
Under what conditions will a newly introduced allele A that is favorable (i.e. selected, so $w_{aa} \leq 1$ and/or $w_{AA} \geq 1$) and present initially in only one copy, survive and fix in the population?

In a finite population without mutation & migration, the favored allele either

- fixes ($p_t \to 1$) with probability $p_0 = \frac{1}{2N}$ (initial frequency), or
- disappears ($p_t \to 0$) with probability $1 - p_0$.

So, the question "Will it fix?" is answered also by the question "Will it be eliminated?" Formally, if X is the extinction event and F is the fixation event, we have $P(X \cup F) = 1$, so $P(F) = 1 - P(X)$.

A Branching Process



Assume:

- All mutant alleles exist in heterozygote genotype, call it h; let Y_h be the random number of *heterozygote mutant offspring* of a heterozygote mutant parent.
- Individuals carrying mutant alleles are *effectively* independent. Why aren't they *actually* independent? Under what conditions are they *effectively* independent?

Branching Process Definition

We need to define the distribution of Y_h. Define $p_0, p_1, \ldots$, where

$$P[Y_h = i] = p_i$$

is the probability that a mutant heterozygote will produce i mutant heterozygote offspring.

We now have that the absolute fitness, or expected number of offspring of individuals of this genotype is

$$\mathrm{E}(Y_h) = \sum_{i=0}^{\infty} i p_i$$

Let λ be the probability that a single mutant A allele (in a single Aa heterozygote, call this fellow the founder) is ultimately lost (goes extinct). It is this quantity we seek to estimate.

Probability of Extinction

If the founder produces no offspring with probability p_0, then extinction happens. If the founder produces 1 offspring, but that offspring begets a lineage that goes extinct (wp λ), then extinction happened. If the founder produces 2 offspring, but both beget a lineage that goes extinct (each wp λ), then extinction happened. And so on...

Therefore, using the Law of Total Probability, extinction from a single starting allele, which happens with probability λ satisfies

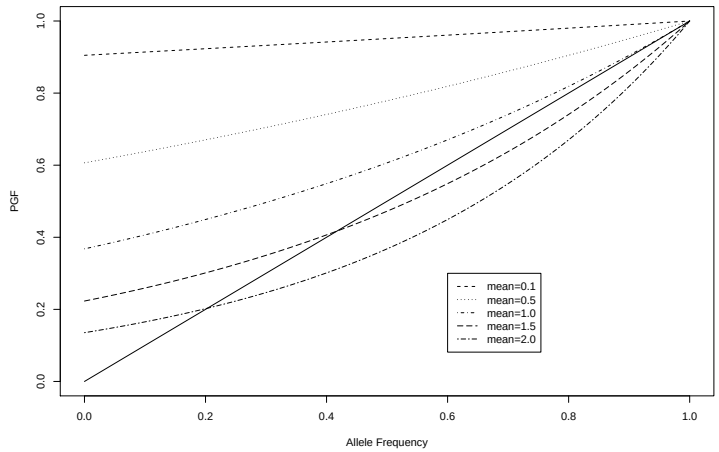
$$\lambda = p_0 + \lambda p_1 + \lambda^2 p_2 + \lambda^3 p_3 + \cdots .$$

This equation may or may not have a solution, other than 1 that is. $\lambda = 1$ clearly always satisfies this equation. There can also be another solution $0 < \lambda < 1$ that is the probability that the newly introduced allele will go extinct. How to find λ?

- Grid search on $\lambda \in (0, 1)$.
- Iteration of $f(\lambda) = p_0 + \lambda p_1 + \lambda^2 p_2 + \cdots$ starting from some initial value.

Example: Poisson Distribution

$$p_0 + \lambda p_1 + \lambda^2 p_2 + \cdots = e^{(z-1)\times \text{mean}}$$



Extinction under Wright-Fisher Model with Selection

So far, our results apply to any distribution p_i of offpsring.

We will now assume a Wright-Fisher model and construct the Wright-Fisher offspring distribution. Suppose again mutant A is the favored allele, and the first individual with the mutation is very likely to be of genotype Aa.

In a Wright-Fisher model, each adult, including our lucky initial Aa, produces an infinite number of gametes. We insert the effects of selection (remember fertility and viability) in our reproduction process.

$$N \text{ adults} \rightarrow \text{meiosis, fertility selection} \rightarrow \text{infinite gamete pool}$$
$$\rightarrow \text{random union} \rightarrow \text{viability selection} \rightarrow \text{pre-adult}$$
$$\rightarrow \text{random kill-off } N \text{ adults}$$

Whether there is a fertility or viability difference, our lucky individual Aa will contribute more to the infinite pool of pre-adults than other individuals. If there is a fertility difference, s/he will contribute more gametes to the gamete pool. If there is a viability difference, his/her offspring will have a better chance of surviving to pre-adulthood. And, of course, there can be combined fertility and viability effects.

To be precise, if the relative fitness of Aa is $1 + s$, then a proportion $\frac{1+s}{N}$ of the pre-adults will be of type Aa and descendent from our lucky individual, which is inflated by s over the usual contribution of parents of $\frac{1}{N}$. Technically, the Wright-Fisher model allows for type AA pre-adults, but there chance of occurrence is very small because in spite of our hero's valient efforts, s/he still contributes very few gametes, relatively speaking, to the infinite gamete pool.

When N of these pre-adults are randomly selected (not by genotype) to survive to adulthood, we have N opportunities to select an Aa type individual. Thus, the number Y_{Aa} of Aa offspring making it into the next generation follows a Binomial distribution. Namely,

$$Y_{Aa} \sim \text{Binomial}(N, \frac{1+s}{N}).$$

When N is large, so the probability of success $\frac{1+s}{N}$ is small and the number of trials is large, the Binomial distribution can be well-approximated by the Poisson distribution with parameter equal to the number of trials

times probability of success. Therefore, we can write instead that

$$Y_{Aa} \sim \text{Poisson}(1+s),$$

so that the pmf is $p_i = \frac{e^{-(1+s)}(1+s)^i}{i!}$.

Probability of Extinction

Now, return to the question of extinction and substitute our new-found p_i into

$$\lambda = p_0 + \lambda p_1 + \lambda^2 p_2 + \cdots .$$

$$\begin{aligned}
\lambda &= e^{-(1+s)} + e^{-(1+s)}(1+s)\lambda + e^{-(1+s)}\frac{(1+s)^2\lambda}{2} \\
&\quad + e^{-(1+s)}\frac{(1+s)^3\lambda}{3!} + \cdots \\
&= e^{-(1+s)}\left[\sum_{i=0}^{\infty}\frac{(1+s)^i\lambda^i}{i!}\right] \\
&= e^{-(1+s)}e^{\lambda(1+s)} = e^{(\lambda-1)(1+s)}.
\end{aligned}$$

Still need numeric solution, but both sides are a tad easier to compute.

Approximate Extinction Probability

If $\lambda \approx 1$, then an approximate solution is available. Then, expanding the Taylor's series, we have

$$\begin{aligned}
\lambda &\approx 1 + (\lambda-1)(1+s) + \frac{(\lambda-1)^2(1+s)^2}{2!} \\
(\lambda-1)&\left[1 - (1+s) - \frac{(\lambda-1)(1+s)^2}{2}\right] \approx 0
\end{aligned}$$

The solution is

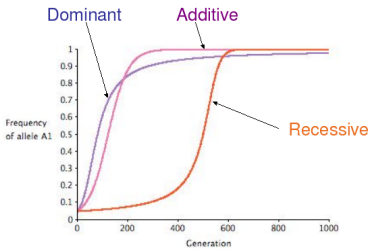
$$1 - \lambda \approx \frac{2s}{(1+s)^2}.$$

- When s is small, then survival of a new mutation $1-\lambda$ is approximately $2s$.
- When s is negative, there is no possibility of survival of the mutant $\lambda = 1$. Why might this prediction *not* be true in real life?

Implications

- When $s = 0.01$, only 1 in 50 new, favorable mutants survive.
- When $s = 0.1$, which is relatively high selection, only 1 in 6 will survive.
- As the number of copies of an allele increases, the chance of its demise decreases. Suppose there are n copies of an allele. If $n \ll N$, the population may be considered large enough so that all mutant lineages can still be considered independent (i.e. all mutants still only mate with nonmutant members of the population). Then, the probability of survival is approximately $1-(1-2s)^n$. If there are 100 copies of the mutant allele and $s = 0.01$, then the probability of survival is 0.86, pretty good.

- Most alleles are lost while they are present in still very low numbers, i.e. in the first few generations after their introduction.



$$\Delta p_A(t+1) = p_A(t)\frac{\bar{w}_A(t)-\bar{w}(t)}{\bar{w}(t)} \quad \text{change due to selection}$$
$$E\left[(\Delta p_A(t+1))^2\right] = \frac{p_A(t)(1-p_A(t))}{2N} \quad \text{change due to drift}$$

- For new mutant alleles, $p_A \approx \frac{1}{2N}$ and $\bar{w}(t) \approx 1$. Under these conditions, genetic drift and selection roughly balance:
  - selection induces change $\Delta p_A \propto \frac{1}{N}$
  - genetic drift induces change $\Delta p_A \propto \frac{1}{N}$
- For more frequent mutant alleles, when $p_A \approx 0.5$, selection has an opportunity to dominate genetic drift
  - *Exercise:* selection induces change $\Delta p_A \propto s$
  - genetic drift induces change $\Delta p_A \propto \frac{1}{\sqrt{N}}$

Weakness of the Approach

Recall that we had to assume that all copies of the mutant allele were independent in the population. To do so, we need to assume a fairly large population and very low mutant allele frequencies.

As the population size declines and the number of mutant alleles increases, it becomes increasingly more likely that two alleles will encounter each other (e.g. mating of Aa and Aa) and will therefore no longer be independent.

The diffusion approximation is proposed to overcome this difficulty and will work for any initial gene frequency.

15.6.3 Diffusion Approximation

Wright Model

N adults $\rightarrow$ meisis, fertility selection $\rightarrow$ infinite gamete pool
$\rightarrow$ random union $\rightarrow$ viability selection $\rightarrow$ pre-adult
$\rightarrow$ random kill-off N adults

Assume no fertility selection and only viability selection. Let the relative viabilities be w_{AA}, w_{Aa}, and w_{aa}.

If there are i copies of mutant A in the parent population, then because there is no fertility selection, there will be $p = \frac{i}{2N}$ A gametes in the gametic pool. The gametes unite at random, so the zygotes pre-viability selection are at HWE with probability of allele A p. After viability selection, the genotype probabilities among pre-adults are

$$p_{AA} = \frac{w_{AA}p^2}{\bar{w}}, \quad p_{Aa} = \frac{2w_{Aa}p(1-p)}{\bar{w}}, \quad p_{aa} = \frac{w_{aa}(1-p)^2}{\bar{w}}$$

The probability of n_{AA}, n_{Aa}, n_{aa} survivors into the next generation is

$$\begin{aligned} P(n_{AA}, n_{Aa}, n_{aa}) &= \binom{N}{n_{AA}\ n_{Aa}\ n_{aa}} \left[\frac{w_{AA}p^2}{\bar{w}}\right]^{n_{AA}} \left[\frac{2w_{Aa}p(1-p)}{\bar{w}}\right]^{n_{Aa}} \\ &\quad \times \left[\frac{w_{aa}(1-p)^2}{\bar{w}}\right]^{n_{aa}} \end{aligned}$$

and the corresponding absolute allele frequencies have probabilities

$$P(j \mid i) = \sum_{k=0}^{j/2} P(k, j-2k, N-j+k)$$

where the arguments are set so $n_{AA} + n_{Aa} + n_{aa} = N$ and $2n_{AA} + n_{Aa} = j$.

Exact Fixation Probabilities

Let u_i be the probability that the mutant allele A ultimately is fixed assuming that it started with i copies. The following equation is true

$$u_i = \sum_j u_j P(j \mid i) \qquad\qquad u = Pu$$

much like our original equation for the extinction probability λ (starting from 1 mutant allele).

We know $u_0 = 0$ since if there are no copies of the mutant allele around, it cannot possibly fix. We also know $u_{2N} = 1$, since if there are $2N$ copies of the allele around it *is* fixed.

With these conditions, there are $2N - 1$ equations (originally $2N + 1$ with 2 fixed) and $2N - 1$ unknowns. One can solve this numerically by computing the $P(j \mid i)$, putting them in a matrix, and then inverting the large matrix. There is no closed-form solution, however.

Under Diffusion Approximation

Let $U(p)$ be the probability that mutant allele A fixes given that it starts at allele frequency p. Specifically $u_i = U(\frac{i}{2N})$.

And, instead of transition probabilities $P(j \mid i)$, define changes in allele frequency, namely let $P_p(\Delta p)$ be the probability that the allele frequency of mutant A changes by Δp given that it is currently p. And, specifically relating back to our previous definition

$$P(j \mid i) = P_{i/2N}\left(\frac{j-i}{2N}\right).$$

We convert $u_i = \sum_j u_j P(j \mid i)$ to the new notation

$$U(p) = \sum_{\Delta p} P_p(\Delta p) U(p + \Delta p),$$

where the sum is over *all* possible changes in allele frequency.

Using Taylor series, we have

$$U(p + \Delta p) \approx U(p) + \Delta p U'(p) + \frac{(\Delta p)^2}{2} U''(p)$$

which we can substitute back into the sum for $U(p)$

$$\begin{aligned} U(p) &\approx \sum_{\Delta p} P_p(\Delta p)U(p) + \sum_{\Delta p} P_p(\Delta p)\Delta p U'(p) \\ &\quad + \frac{1}{2}\sum_{\Delta p} P_p(\Delta p)(\Delta p)^2 U''(p) \\ &= U(p)\sum_{\Delta p} P_p(\Delta p) + U'(p)\sum_{\Delta p} P_p(\Delta p)\Delta p \\ &\quad + \frac{U''(p)}{2}\sum_{\Delta p} P_p(\Delta p)(\Delta p)^2 \\ &= U(p) + U'(p)\mathrm{E}(\Delta p) + \frac{U''(p)}{2}\mathrm{E}\left[(\Delta p)^2\right]. \end{aligned}$$

Rearrange to obtain

$$\mathrm{E}(\Delta p)U'(p) + \frac{U''(p)}{2}\mathrm{E}\left[(\Delta p)^2\right] \approx 0.$$

Diffusion Assumptions

Assumes that $P_p(\Delta p)$ is small except when Δ_p is small. In other words, it assumes that only small changes are likely in each generation. But this is the same as assuming that population sizes are large and selection coefficients are not large, for either of these could change the population allele frequencies very fast (the first by random chance) the second by a strong deterministic force.

The solution (see Felsenstein's *Theoretical Evolutionary Genetics* for details) is

$$U(p) = \frac{\int_0^p G(x)dx}{\int_0^1 G(x)dx},$$

where

$$G(x) = \exp\left[-2\int_c^x \frac{E(\Delta p)}{E\left[(\Delta p)^2\right]}dp\right].$$

The c cancels in the ratio, so it need not be specified.

Let $M(p) = \mathrm{E}[\Delta p]$ and $V(p) = \mathrm{E}\left[(\Delta p)^2\right]$. The exact form of these depends on the type of selection operating.

Multiplicative Selection

We start with multiplicative selection, so

AA	Aa	aa
$(1+s)^2$	$1+s$	1

Our sequence of events (as a reminder) is

N adults $\rightarrow$ meiosis, fertility selection $\rightarrow$ infinite gamete pool
$\rightarrow$ random union $\rightarrow$ viability selection $\rightarrow$ pre-adult
$\rightarrow$ random kill-off to N adults

Selection acts on the infinite pool (of gametes or pre-adults), so it is deterministic. We can resort back to our early work and recall that for this case

$$\Delta p = p_{t+1} - p_t = \frac{sp_t(1-p_t)}{1+sp_t}.$$

$M(p)$

The mean change in allele frequency is equal to the above deterministic quantity, so

$$M(p) = \frac{sp(1-p)}{1+sp},$$

where I have dropped the dependence on time t. The diffusion approximation applies when s is small enough (and also N is large enough) that huge changes Δp are not expected, so an approximate formula under these conditions is

$$M(p) \approx sp(1-p).$$

$V(p)$

Now it is time to select N random pre-adults to survive to adulthood.

Selection of N survivors is *not* the same as random sampling of $2N$ gametes from a pool with frequency p (Wright-Fisher model without selection). There are two problems:

- **Problem 1.** The allele frequency has changed by Δp, and
- **Problem 2.** Selection at the genotype level can introduce dependence between the alleles sampled from the same individual.

What is the new allele frequency?

Let $P'_{AA}, P'_{Aa}, P'_{aa}$ be the genotype frequencies after selection in the pre-adult pool.

Let $p'_A = P'_{AA} + \frac{1}{2}P'_{Aa}$ be the corresponding allele frequency in the pre-adult pool after selection.

The genotype frequencies in the pre-adults will be

$$\begin{aligned} P'_{AA} &= \frac{w_{AA}p^2}{\bar{w}} = \frac{(1+s)^2p^2}{\bar{w}} \\ P'_{Aa} &= \frac{2w_{Aa}p(1-p)}{\bar{w}} = \frac{2(1+s)p(1-p)}{\bar{w}}, \end{aligned}$$

where

$$\bar{w} = (1+sp)^2.$$

Then, the mutant allele frequency among the pre-adults will be

$$p'_A = P'_{AA} + \frac{1}{2}P'_{Aa} = \frac{(1+s)^2p^2 + (1+s)p(1-p)}{(1+sp)^2} = \frac{(1+s)p}{1+sp}.$$

(*Exercise*: Verify when s is small, $p'_A \approx p_A$, but we already knew this.)

Compare

$$\begin{aligned} P'_{AA} &= \frac{(1+s)^2p^2}{(1+sp)^2} \\ p'_A &= \frac{(1+s)p}{1+sp}. \end{aligned}$$

What do you observe?

By making the diffusion approximation we have assumed Δp is small. The first problem can be approximated away.

The second problem is no problem when we have multiplicative selection.

So now we can argue that selecting N pre-adults to make it to adulthood is like selecting $2N$ alleles at random from a pool where A is present in frequency approximately p.

Selecting $2N$ alleles from a pool with a proportion $\sim p_t$ of A alleles is binomial sampling, and the variance in allele frequency after sampling is

$$\text{Var}(p_{t+1} \mid p_t) \approx \frac{p_t(1-p_t)}{2N}.$$

We have also argued previously that $\text{Var}(\Delta p_{t+1}) = \text{Var}(p_{t+1} \mid p_t)$.

Dropping the dependence on t and recalling that $V(p) = \text{E}\left[(\Delta p)^2\right] = \text{Var}(\Delta p) + \{\text{E}[\Delta p]\}^2$, we conclude that

$$V(p) \approx \frac{p(1-p)}{2N} + M^2(p) \approx \frac{p(1-p)}{2N},$$

the final approximation made because of our assumption that allele frequency change is small, therefore mean allele change is small, especially when squared.

$U(p)$

Recall

$$G(x) = \exp\left[-\int_c^x \frac{2M(p)}{V(p)}dp\right].$$

Here,

$$\frac{2M(p)}{V(p)} \approx 2sp(1-p)\frac{2N}{p(1-p)} = 4Ns.$$

Integrating this yields

$$\int_c^x 4Nsdp = 4Ns(x-c).$$

The resulting fixation probability is

$$\begin{aligned} U(p) &= \frac{\int_0^p G(x)dx}{\int_0^1 G(x)dx} \\ &\approx \frac{\int_0^p e^{-4Ns(x-c)}dx}{\int_0^1 e^{-4Ns(1-c)}dx} = \frac{\int_0^p e^{-4Nsx}dx}{\int_0^1 e^{-4Ns}dx} = \frac{1-e^{-4Nsp}}{1-e^{-4Ns}}. \end{aligned}$$

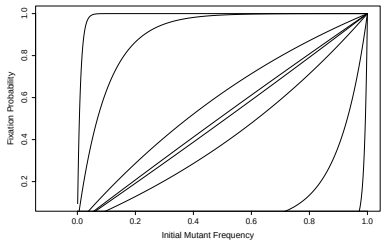
Strength of Approximation

	$s = 0.01$		$s = 0.1$	
p	exact	approx.	exact	approx.
0.05	0.06002	0.06006	0.17873	0.18465
0.1	0.11885	0.11894	0.32602	0.33583
0.2	0.23305	0.23321	0.54756	0.56095
0.3	0.34279	0.34300	0.69830	0.71184
0.4	0.44825	0.44848	0.80100	0.81299
0.5	0.54959	0.54983	0.87107	0.88080
0.7	0.74056	0.74077	0.95166	0.95761

$N = 10$

The diffusion model assumes allele frequency changes are tiny, but well approximates, as far as ultimate outcome, the actual, more jagged changes.

Implications/Interpretation



When $s = 0$, fixation probability is initial frequency.

When $s > 0$, fixation probability is increased. In figure, $4Ns$ varies from -100 to 100. For $4Ns = 0.1$ (curve just north of $y = x$), there is hardly any improvement, but for $4Ns = 100$ (top curve) fixation is virtually certain for any non-trivial initial frequency.

Rule of Thumb Natural selection is effective against drift if one individual dies because of genetic causes every two generations.

Replace N with N_e when dealing with non-Wright-Fisher populations.

Weak Selection

We understand that selection is a powerful force when $|4Ns| > 1$, but what happens when $4Ns$ is small. To consider this, we return to our solution for $U(p)$

$$U(p) \approx \frac{1 - e^{-4Nsp}}{1 - e^{-4Ns}},$$

and expand the numerator and denominator in Taylor's series

$$\begin{aligned} U(p) &\approx \frac{1 - \left[1 - 4Nsp + \frac{1}{2}(-4Nsp)^2\right]}{1 - \left[1 - 4Ns + \frac{1}{2}(-4Ns)^2\right]} \\ &= \frac{4Nsp - 8Ns(Nsp^2)}{4Ns - 8Ns(Ns)} \\ &= \frac{p - 2Nsp^2}{1 - 2Ns} \end{aligned}$$

When $4Ns$ and therefore $2Ns$ is small, then $U(p) \approx p + 2Nsp(1 - p)$.

Dominance

Consider the general dominance scheme

AA	Aa	aa
$1 + s$	$1 + hs$	1

with h measuring the degree of dominance of A over a.

From previous results and assuming s is small so s^2 terms can be dropped, we have the average allele frequency change is

$$M(p) \approx sp(1 - p)\left[p(1 - sh) + h\right].$$

We also set

$$V(p) \approx \frac{p(1-p)}{2N_e}$$

to yield

$$G(x) = e^{-2N_e s(1-2h)x^2 - 4N_e shx}.$$

To get $U(p)$ requires integrating $G(x)$ and there is no explicit integral of $G(x)$, but it can be integrated numerically to obtain the following results.

- When $h \approx 1$, the mutant allele A is dominant. When p is small and when $4N_e s$ is large, this behaves like the multiplicative case. The only thing that matters is the fitness of the heterozygote, i.e. hs is all that matters.
- When $h = 0$, the mutant allele A is recessive. A rare recessive allele experiences no selection. It will have a very low chance of fixing.
- When $h > 1$, the mutant allele A shows overdominance. In infinite populations there would be an equilibrium established between the mutant and non-mutant alleles. In finite populations, it has a fair chance of being carried to higher frequency. Once at high enough frequency it will approach its equilibrium. It will stay at equilibrium with random excursions away, but eventually in one of those excursions, it will fix or be lost by chance events.

Discussion

Can you describe a finite population situation where fixation will not occur?

Equilibrium Distribution

We now consider finite population size, selection, and mutation/migration *simultaneously*.

Specifically, we consider what happens after the population has proceeded for many, many generations under the same conditions.

In infinite populations, an equilibrium allele frequency is eventually obtained when the underlying deterministic forces are balanced. The same equilibrium is approached in finite populations, but because of genetic drift, there are random fluctuations around the equilibrium allele frequency. What *fixes* is the probability distribution that describes this population. Suppose X_t is the number of A alleles at time t. Then as $t \to \infty$, X_t will not approach some equilibrium value, but the distribution that describes X_t will approach some fixed distribution:

$$P(X_t = i) \to f_i$$

constant for all time t after equilibrium is achieved.

At equilibrium, we can imagine multiple replicate populations and describe the distribution of allele frequencies across these populations using f_i. What do you expect is the value of $E[X_t]$ at equilibrium?

We seek the transition probability $P(j \mid i)$ that the population has j alleles after one generation of reproduction, selection, mutation/migration, having been at i alleles one generation before.

- Compute the starting gene frequency $p = \frac{i}{2N}$.
- The proportion of A in the initial infinite gamete pool is p.

- A fraction u of the A gametes mutate to a. A fraction v of the a gametes mutate to A. The new allele frequency after mutation will be
$$p' = (1-u)p + v(1-p).$$

- A fraction m of the gametes are replaced by immigrants with allele A frequency p_I. After migration, the allele frequency is
$$p^* = (1-m)p' + mp_I.$$

- Assume random union of gametes and compute HWE genotype frequencies.

- Apply selection and renormalize genotype probabilities. Let these new genotype proportions be P, Q, and R.

- Compute the trinomial probability for all possible combinations of N surviving adults
$$P(k, l, N-k-l) = \frac{N!}{k!l!(N-k-l)!} P^k Q^l R^{N-k-l}.$$
When selection is multiplicative, sampling N adults is equivalent to sampling $2N$ genes. We used this before.

- The probability that there are j A genes among these N adults is the sum of all probabilities above such that $2k + l = j$.

Let the equilibrium distribution be represented by f_i, where f_i is the probability that a random population will have i mutant A alleles at equilibrium. The equilibrium is found as the solution of

$$\begin{aligned} P(X_{t+1} = j) = f_j &= \sum_{i=0}^{2N} P(X_t = i, X_{t+1} = j) \\ &= \sum_{i=0}^{2N} P(X_{t+1} = j \mid X_t = i) P(X_t = i) \\ &= \sum_{i=0}^{2N} P(j \mid i) f_i. \end{aligned}$$

if the population is at equilibrium at time t and using the Law of Total Probability. There are $2N+1$ equations for $j = 0, 1, \ldots, 2N$, but $\sum_j f_j = 1$ leaving $2N$ equations to solve.

These equations can be solved numerically if N is sufficiently small.

Diffusion Approximation

Replace f_i by $f(p)$ and $P(j \mid i)$ by $P_p(\Delta p)$ so we are following gene frequencies instead of absolute counts. The equations to solve become
$$f(p) = \sum_{\Delta p} f(p - \Delta p) P_{p-\Delta p}(\Delta p).$$

Replace $f(p)$, which is discrete for $p = 0, \frac{1}{2N}, \frac{2}{2N}, \ldots$ by a continuous approximation
$$f(p) = \phi(p) dp,$$

and plug back into the equation to obtain
$$\phi(p) dp \approx \sum_{\Delta p} \phi(p - \Delta p) P_{p-\Delta p}(\Delta p) dp.$$

$\phi(p)$ is the continuous approximation to the equilibrium distribution that is our main interest.

Then, we use Taylor series expansions of three terms around p to replace $\phi(p-\Delta p)$ and $P_{p-\Delta p}$

$$\begin{aligned}\phi(p)dp &\approx \sum_{\Delta p}\left[\phi(p)-\Delta p\phi'(p)+\frac{(\Delta p)^2}{2}\phi''(p)\right] \\ &\quad\times\left[P_p(\Delta p)-\Delta pP_p'(p)+\frac{(\Delta p)^2}{2}P_p''(\Delta p)\right]\delta p.\end{aligned}$$

Rearrange and ignore terms involving $(\Delta p)^3$ or $(\Delta p)^4$. Also, define

$$\begin{aligned}M(p) &= \mathrm{E}(\Delta p)=\sum_{\Delta p}P_p(\Delta p)\Delta p \\ V(p) &= \mathrm{E}[(\Delta p)^2]=\sum_{\Delta p}P_p(\Delta p)(\Delta p)^2\end{aligned}$$

With these definitions, we also have

$$\begin{aligned}M'(p) &= \frac{dM(p)}{dp}=\sum_{\Delta p}P_p'(p)\Delta p \\ V'(p) &= \frac{dV(p)}{dp}=\sum_{\Delta p}P_p'(p)(\Delta p)^2 \\ V''(p) &= \frac{d^2V(p)}{dp^2}=\sum_{\Delta p}P_p''(p)(\Delta p)^2.\end{aligned}$$

The final equation obtained is

$$0\approx-\frac{d}{dp}\left[M(p)\phi(p)\right]+\frac{1}{2}\frac{d^2}{dp^2}\left[V(p)\phi(p)\right].$$

Solving the derived differential equation is not pretty and will not be further discussed. The solution is

$$\phi(p)=\frac{K}{V(p)}\exp\left[2\int_c^p\frac{M(x)}{V(x)}dx\right],$$

where K is fixed so that the density function $\phi(p)$ integrates to 1 over the interval $[0,1]$. c is also part of the constant, but can be set for convenience and then K appropriately adjusted.

Our focus now is to determine $M(p)$ and $V(p)$ for various cases of interest.

Mutation and Drift

Suppose allele A can mutate to a with probability u and a can mutate to A with probability v.

From infinite populations, the expected change in allele frequency is

$$M(p)=\mathrm{E}(\Delta p)=-up+v(1-p).$$

As for $V(p)$, if we assume conditions where (or compute N_e such that) selecting N individuals for survival into the next generation is equivalent to randomly selecting alleles, then we have

$$V(p)\approx\frac{p(1-p)}{2N_e},$$

where, again, we assume we can ignore $M^2(p)$, and the allele frequency in the infinite gamete pool is approximately p, even though mutation has slightly altered it. These approximations are justified when u, v are small and N_e is quite large.

Plug these $M(p)$ and $V(p)$ into our solution for $\phi(p)$ to find

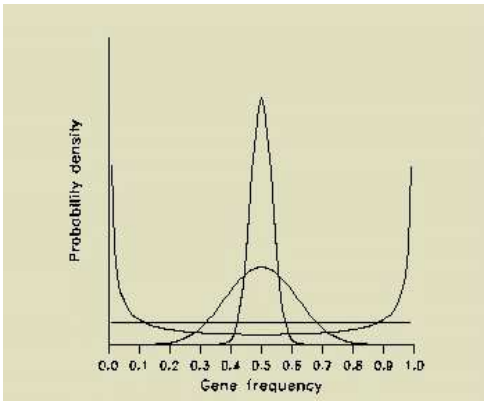
$$\phi(p) = Kp^{4N_e v-1}(1-p)^{4N_e u-1}.$$

You may recognize this as the Beta distribution. Basic properties:

$$\begin{aligned} \mathrm{E}(p) &= \frac{v}{u+v} \\ \mathrm{Var}(p) &= \frac{\bar{p}(1-\bar{p})}{1+4N_e u+4N_e v}, \end{aligned}$$

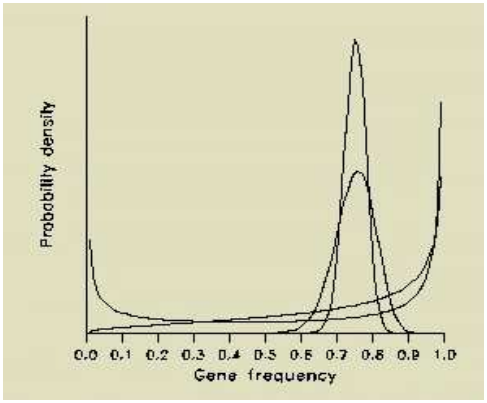
where $\bar{p} = \mathrm{E}(p)$.

Balanced Mutation and Drift



Here the forward and backward mutation rates are equal $u = v$. $4N_e u$ varies through $100, 10, 1, 0.1$. When $4N_e u < 1$, the distribution becomes U-shaped and the populaton tends to spend much time with one allele fixed.

Unbalanced Mutation and Drift



Forward mutation rate is larger $u = 3v$. $4N_e u$ varies through $50, 15, 0.5, 0.1$.

Migration and Drift

An equilibrium is established between migration and drift when immigrants come in with a constant allele frequency, and prevent fixation on the island.

The expected change in allele frequency is

$$M(p) = mQ + (1-m)p - p = m(Q-p),$$

when the allele frequency of immigrants is Q and in the home population is p.

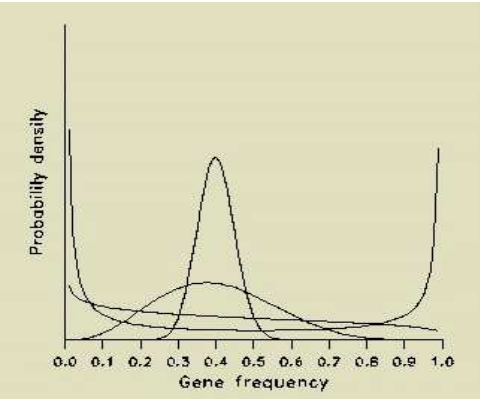
We approximate $V(p)$ as usual by assuming m is small and N is large

$$V(p) \approx \frac{p(1-p)}{2N_e}.$$

This is essentially equivalent to the mutation model and the equilibrium distribution is also Beta

$$\phi(p) = Kp^{4N_e mQ-1}(1-p)^{4N_e m(1-Q)-1}.$$

Unbalanced Migration and Drift



$Q = 0.4$ is the allele frequency in migrants. $4N_e m$ varies through $100, 10, 2, 0.1$.

General Selection vs. Drift

For the general case (no specific structure imposed on the relative fitnesses) when relative mean fitness of the population is $\bar{w} = w_{AA}p^2 + 2w_{Aa}p(1-p) + w_{aa}(1-p)^2$, we can think of the mean relative fitness as a function of $\bar{w}(p)$ of p.

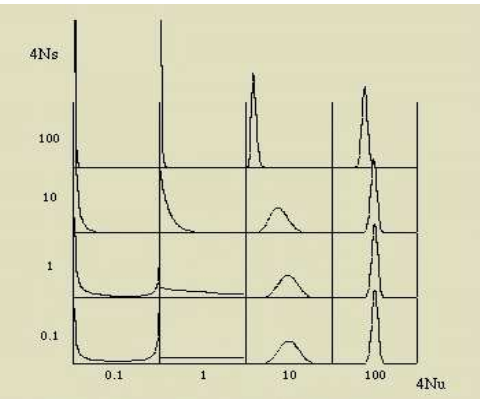
The equilibrium distribution for this general case is

$$\phi(p) = Kp^{4N_e v-1}(1-p)^{4N_e u-1}\left[\bar{w}(p)\right]^{2N_e}.$$

This is like the Beta distribution except that it is inflated near the peaks of $\bar{w}(p)$ because it is multiplied by $\bar{w}(p)$ raised to a fairly large value $2N_e$.

So, mutation draws the allele frequency toward the mutation equilibrium and selection draws the allele frequency toward the frequency which leads to the largest mean population fitness.

Selection and Drift



Multiplicative selection with allele A selected over allele a. $4Ns$ varies through $100, 10, 1, 0.1$ from top to bottom. $4Nu$ varies through $0.1, 1, 10, 100$ from left to right.