Natural Selection and Genetic Drift: An Exploration of Allele Frequencies Within a Population

Leslie Mallinger

8 May 2008

1 Background

An allele is one of two or more alternative forms of a gene that arise by mutation and are found at the same place on a chromosome. Because each person's genome contains two copies of each chromosome (one from his father and one from his mother), each individual has two alleles for each gene. These alleles may be the same or they may be different from one another (Figure 1). In this study, we use mathematical modeling to simulate and investigate the effects of natural selection and genetic drift on the frequency of alleles within a population.

For simplicity, we will solely consider an allele which is present in the population in only two forms, form A and form B. In this case, there are three possible allele combinations, since each individual has two alleles: AA, AB, and BB. Allele combinations where both of an individual's alleles are of the same type are termed *homozygous*, whereas the allele combination AB is termed *heterozygous*. These allele combinations make up an individual's *genotype*, which combines with environmental effects to cause the observable traits or characteristics that make up the individual's *phenotype*.

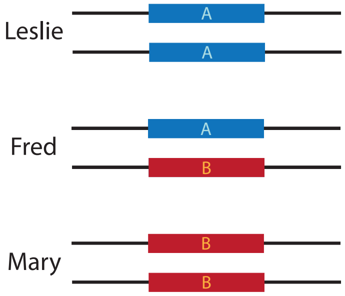


Figure 1: *The three possible genotypes for individuals within a population which contains two forms of a given allele. Leslie has two copies of the A allele, so her genotype is AA and she is said to be "homozygous A." Fred has one copy of the A allele and one copy of the B allele, so his genotype is AB and he is said to be "heterozygous." Mary has two copies of the B allele, so her genotype is BB and she is "homozygous B."*

2 Genetic Drift

Changes in allele frequencies are one indicator of evolution within a population. Genetic drift is the force that acts to change allele frequency simply based on sampling error during mating. In a theoretical population where the population size remains constant, all individuals survive to reproductive age and reproduce with the same frequency, mating is random, and there are no occurrences of migration or genetic mutation, genetic drift is the only evolutionary force at work.

To conceptualize genetic drift, consider an isolated population of ten individuals (Figure 2), Generation 1). Of the twenty alleles that these individuals contribute to the gene pool, ten of them are A alleles, and ten are B alleles—allele A is present in this generation at a frequency of 50%, or 0.5. We simulate random mating by randomly choosing an allele from the gene pool and designating a new allele in the next generation to be of the same type. The chosen allele is replaced in generation 1's gene pool, and the selection process is repeated until generation 2's gene pool contains twenty alleles (ten individuals). Because of the nature of random sampling, it is highly possible—particularly in small populations such as this one—that the number of each allele type will shift between generations. In this case, sampling error caused generation 2 to inherit eight A alleles and twelve B alleles rather than the expected ten alleles of each type. Allele A is present in generation 2 at a frequency of 0.4[1].

If we imagine this process repeating over the course of several generations, we can see that the number of A and B alleles will continue to fluctuate until one or the other is no longer present in the population. Because that allele is no longer in the gene pool and we are assuming that there is no method through which it can be reintroduced, the population

[1] Note that the frequency of allele A and the frequency of allele B sum to 1, so determination of only one allele type is sufficient to represent the state of the population.

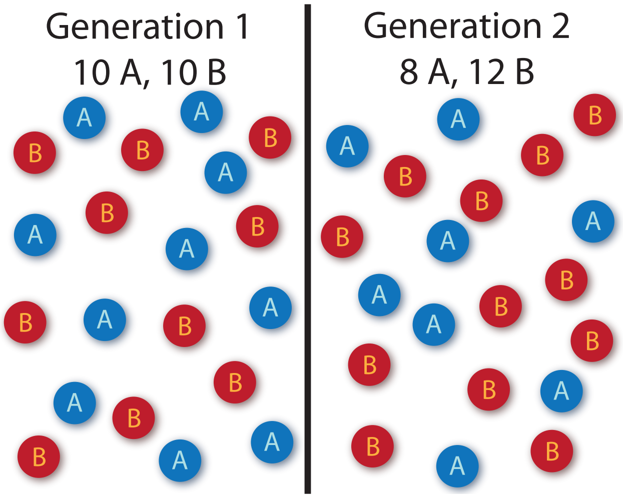


Figure 2: *A visual representation of genetic drift in a small population. The gene pool for generation 1 contains ten A alleles and ten B alleles, but sampling error during mating causes the allele distribution to drift towards an abundance of B alleles in the next generation.*

has come to an equilibrium with only one allele. The allele that has disappeared from the population is said to have gone to *deletion*, whereas the allele that is still present has gone to *fixation*. It can be shown that given enough time, genetic drift will cause all alleles to become either fixed or deleted in a population, and that a given allele's likelihood of going to fixation is equivalent to its initial frequency. Because sampling error is more pronounced when smaller sample sizes are used, genetic drift causes more variation in frequency in smaller populations.

2.1 Modeling Genetic Drift

We investigate two different methods by which to model the phenomenon of genetic drift. The first, "brute force," method is to simulate a gene pool and to choose randomly from that gene pool to create the subsequent generation, much as we did in the example above. The second method utilizes the binomial distribution, which is the discrete probability distribution of the number of successes in a sequence of independent yes/no experiments with a given probability of success. We compare the two methods using a variety of criteria to determine which is better suited to the task at hand. All modeling was performed using MATLAB.

For consistency, we use the following terminology:

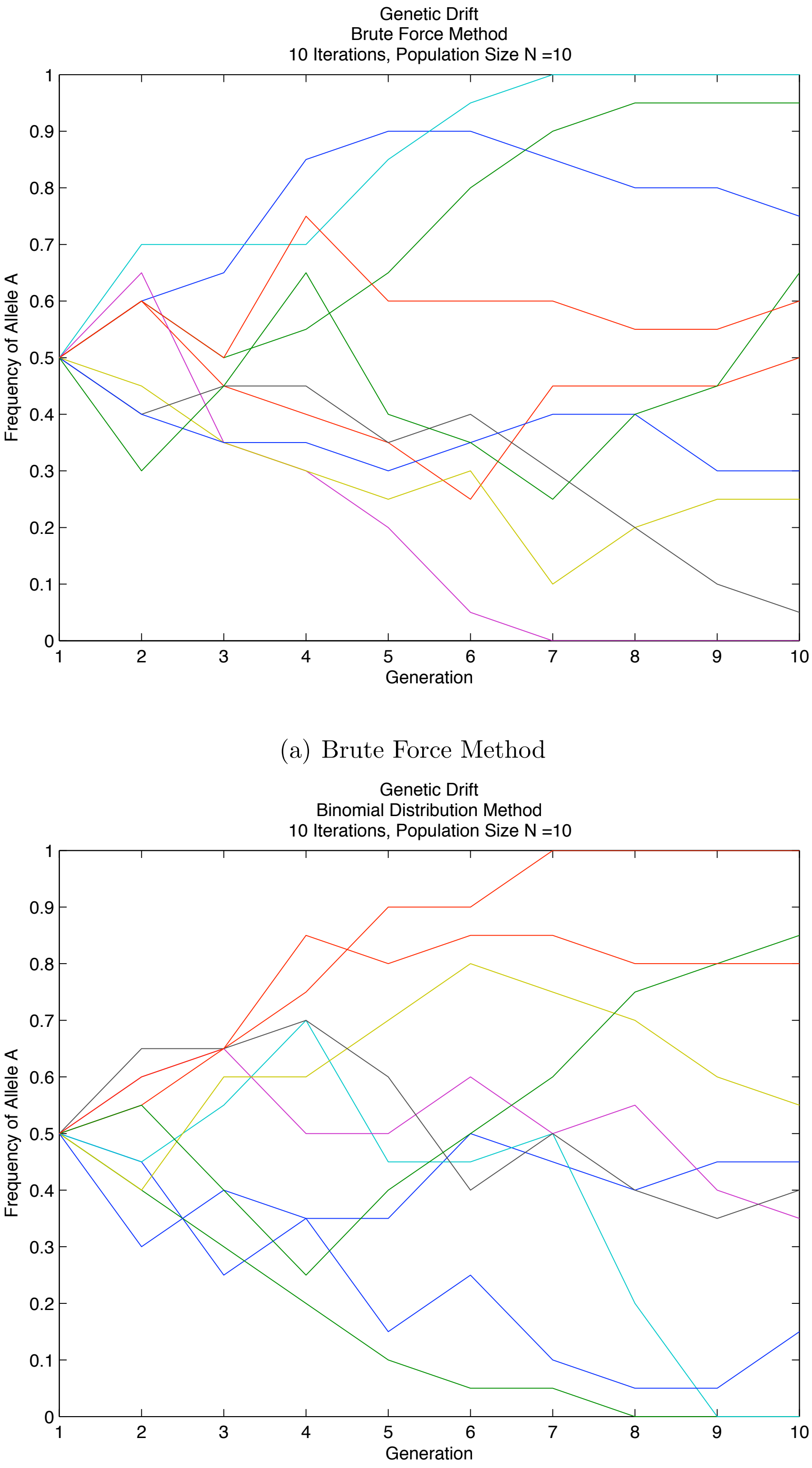
- N = number of individuals within the population (population size)
- g = number of generations for which the simulation is run
- a = initial allele frequency of allele A (0 to 1)

2.1.1 Brute Force Method

To simulate the random picking of gametes from a gene pool, we create a vector of "alleles" which consists of $2N$ random real numbers on the unit interval. If the number is less than a, that entry corresponds to an A allele and its value is reset to 1. All other entries correspond to B alleles and their values are reset to 0. Summing the components of the vector gives the number of A alleles in the new population, and dividing this number by $2N$, the total number of alleles, gives a value for the allele frequency of the subsequent generation. This process is repeated for the desired number of generations. The frequency of allele A is plotted as a function of generation (Figure 3(a)).

2.1.2 Binomial Distribution Method

Instead of literally simulating the choosing of gametes, the binomial distribution method uses a discrete probability distribution to determine the allele frequencies in each generation. We utilize MATLAB's `binornd(n,p)` function to generate random numbers from the binomial distribution with $n = 2N$ gametes and $p =$ the frequency of allele A in the previous generation. The output of the function is then taken to be the allele frequency of the next generation. As with the brute force method, this process is repeated for the desired number of generations and allele frequency is plotted as a function of generation (Figure 3(b)).



(a) Brute Force Method

(b) Binomial Distribution Method

Figure 3: *Sample graphs for each of the methods of modeling genetic drift. In these examples, population size is $N = 10$, allele A has an initial frequency of $a = 0.5$, and 10 iterations are run over $g = 10$ generations each.*

2.1.3 Comparing the Two Methods

To compare the two methods, we first perform a visual comparison of the plots that they generate. In Figure 3 we notice that although the graphs are not identical—this is expected because genetic drift is a stochastic process—there are no visually discernible differences between the two methods of simulation. This is true of the numerous graphs that were generated over the course of this project.

We contrast the efficacy of the methods by examining their runtimes. Figure 4 shows the results of performing six hundred trial runs with each method. The six trial types were increasingly difficult, with one hundred iterations per trial per method. It is apparent that the brute force method is more efficient than the binomial distribution method. MATLAB profiler reveals that calling the `binornd` function is the slowest step in the binomial distribution method[2].

The final technique we use to compare the brute force and binomial distribution methods is the comparison of the average number of generations until the simulated population's allele frequency reaches deletion or fixation. The results of this analysis revealed no significant difference between the two methods and will be discussed further in section 2.2.

[2] Interestingly enough, `http://web.mit.edu/mskilic/www/MyStatsFunctions/BinoRnd/main.html` indicates that `binornd` uses the same method of number generation as was implemented in the brute force method described above. The prominent difference in runtimes for the two methods implies that my implementation is somehow faster than `binornd`'s.

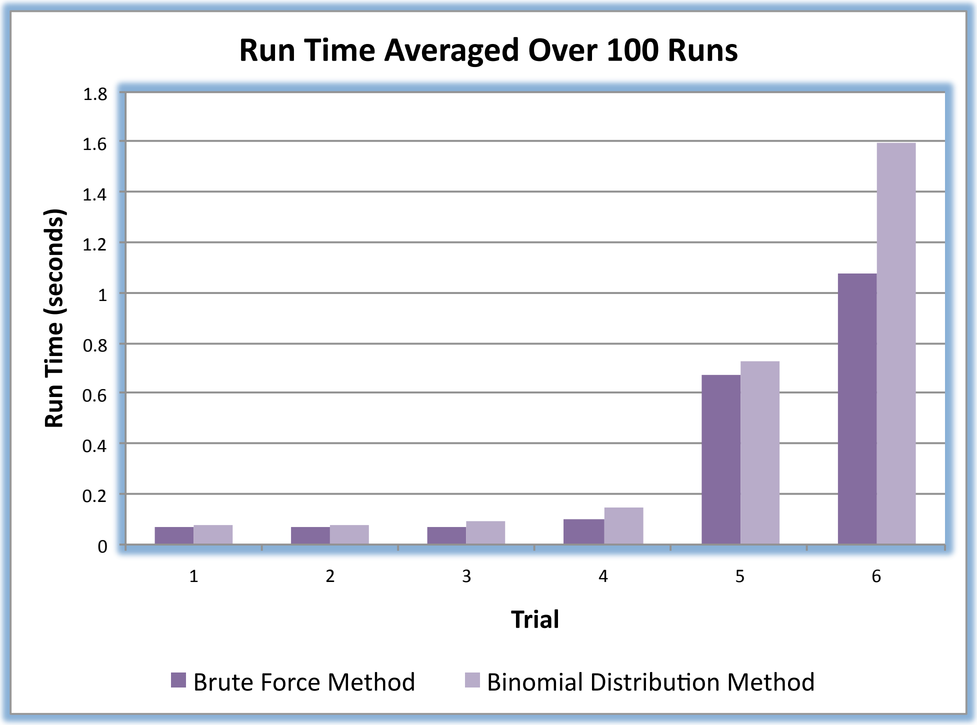


Figure 4: *The run time averaged over one hundred runs for each of six different and increasingly difficult trials using the brute force and binomial distribution methods of modeling genetic drift. The brute force method is demonstrably faster than the binomial distribution method.*

2.2 Investigating Trends in Genetic Drift

We first explore the effects of changing population size on the number of generations until an allele reaches deletion or fixation. For simplicity, we set $a = 0.5$ so that both allele A and allele B have an equal probability of going to either extreme. Figure 5 shows sample plots for populations with ten, one hundred, and one thousand individuals. As expected, there is more pronounced variation in allele frequency due to genetic drift in the smaller populations than in the larger ones. Within fifty generations, all ten iterations for the population with ten individuals led to either fixation or deletion. It took nearly one hundred generations for the iterations for the population with one hundred individuals to even begin to reach that state. In the largest population, genetic drift produced little enough effect that the allele frequency did not stray more than 20% in any direction within the allotted time.

For a quantitative analysis of this phenomenon, we record the average number of generations until deletion over one thousand runs for a variety of population sizes (Figure 6)[3]. We find that average generations until deletion appears to increase linearly with population size. This figure also illustrates that neither the brute force method nor the binomial distribution method of modeling results in a consistently higher number of generations, so we conclude that the two methods have equivalent functionality for our purposes.

Next we explore the effects of changing allele frequency. Figure 7 illustrates that, as expected, an allele's likelihood of going to fixation is equivalent to its initial frequency. For example, an allele with initial frequency 0.1 went to fixation 104 out of 1,000 times that the brute force method was used to simulate drift. The allele went to deletion the remaining 896 times. The slight deviation from the expected 100:900 ratio is because genetic drift is a stochastic process; if we were to sample over more than 1,000 runs, we would likely find more accurate distributions.

Figure 7 also suggests that the brute force method routinely produced fewer deletions than expected. To test for the presence of this bias, the brute force method was run 20,000 times with an initial allele frequency of 0.5. This iteration resulted in 0.46% more deletions than fixations, so we can conclude that the brute force method does not systematically produce more deletions than expected, and that the simulation is in fact relatively balanced. Similar analysis of the binomial distribution method reveals that it, too, is evenly balanced.

[3] As expected, the plot of average number of generations until fixation is nearly identical to Figure 6.

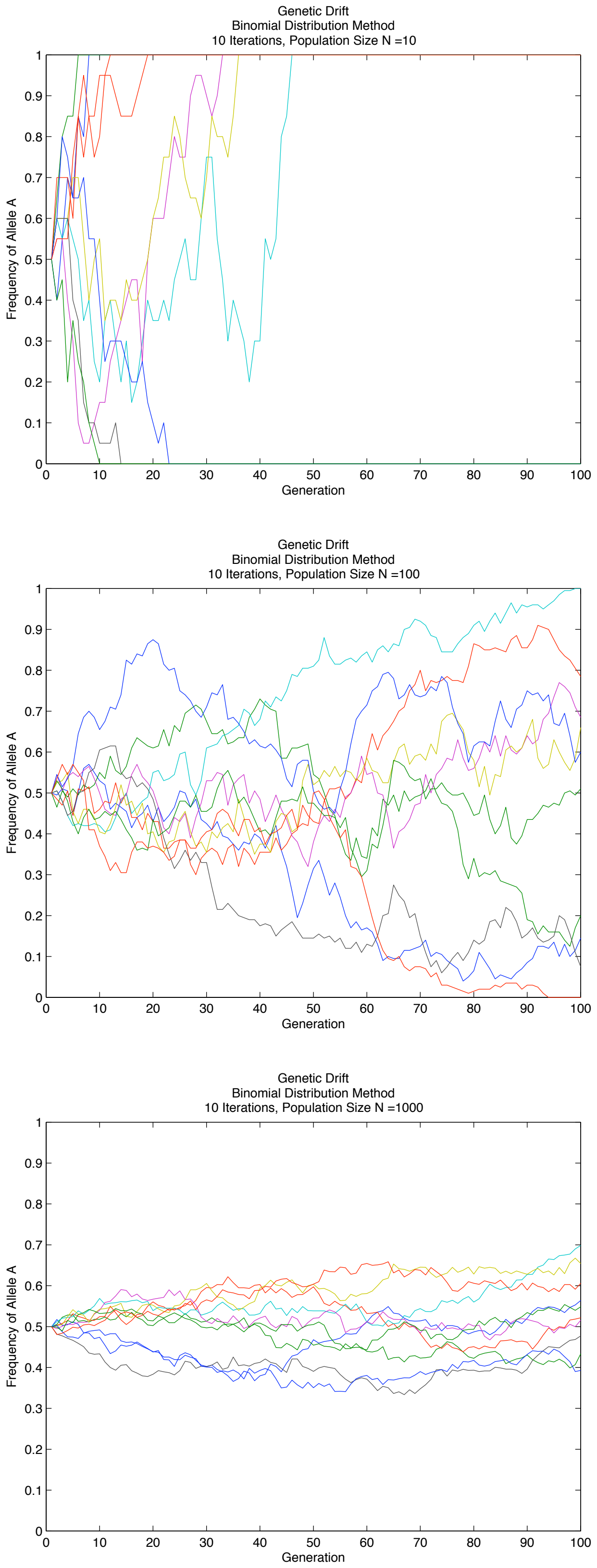


Figure 5: *Plots illustrating the trend in genetic drift as population size increases. The uppermost graph simulates $N = 10$, the middle graph $N = 100$, and the lowermost graph $N = 1000$.*

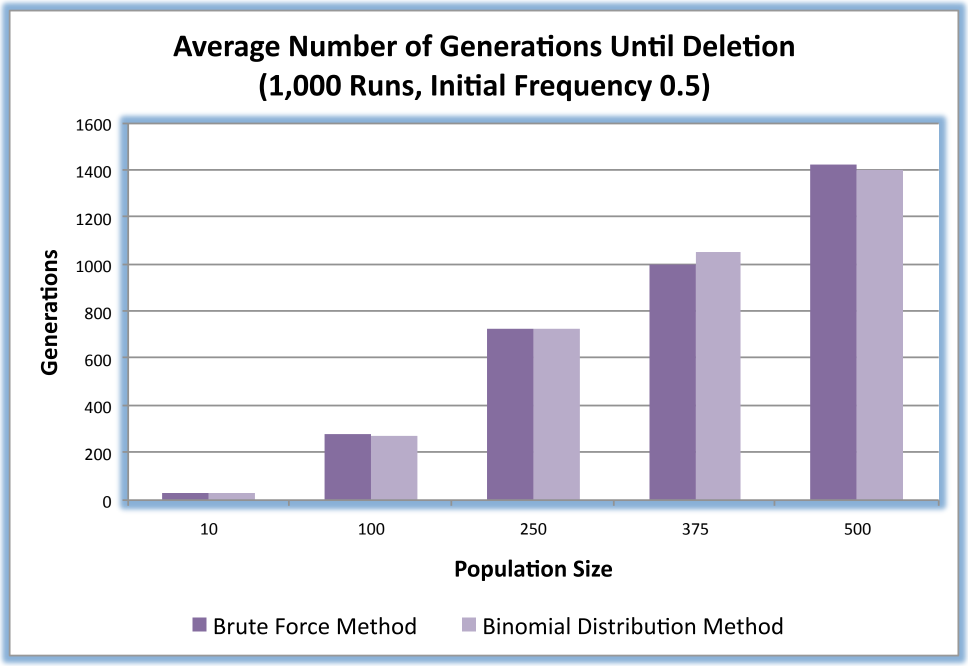


Figure 6: *Average number of generations until deletion over one thousand runs for a variety of population sizes. Generations until deletion increases approximately linearly with population size.*

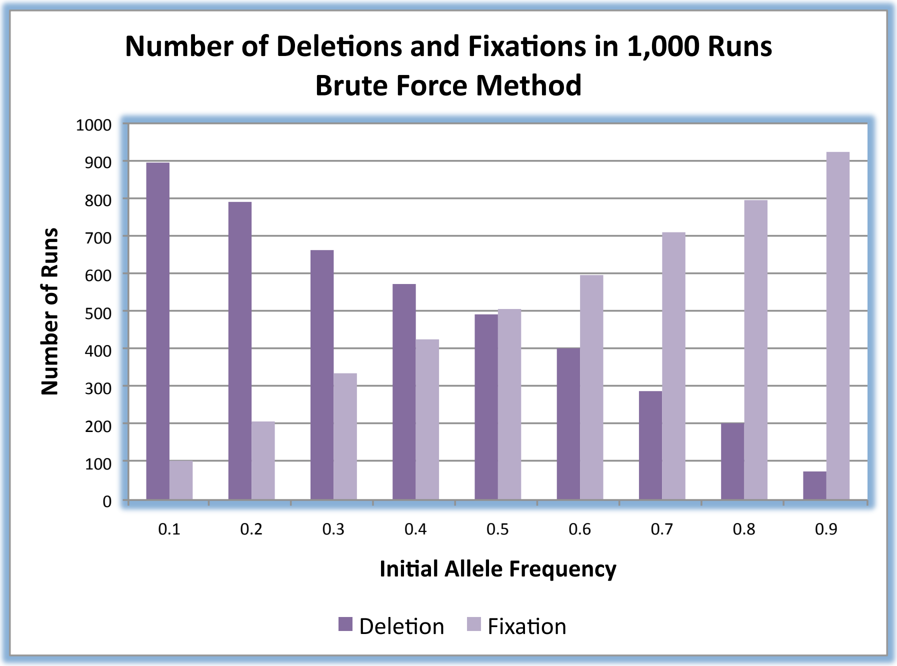


Figure 7: *Number of deletions and fixations in 1,000 runs of the brute force method of modeling genetic drift. The probability that an allele will go to fixation is effectively equivalent to its initial frequency.*

3 Natural Selection

Natural selection is another mechanism which can cause changes in a population's allele frequencies. It is based on the fact that individuals with particular phenotypes—observable traits or characteristics—have a higher likelihood of surviving to reproductive age and/or producing more offspring than do individuals with other phenotypes. When this difference in reproductive success can be passed on from generation to generation, natural selection can lead to evolution.

Consider an individual with genotype AA. If this individual has such traits that it is more likely to produce offspring than are AB or BB individuals, then we say that it has a higher *fitness*. Fitness is given as a number in the range $[0.0, 1.0]$, where the two extremes indicate that an individual will either never or always reproduce, respectively.

In this paper, we assume that phenotypic traits are determined only by genotype and environmental factors do not influence an individual's fitness. We also assume that genetic drift is not at work in this population; that is, sampling error does not occur, so eggs and sperm in the gene pool collide with each other at their actual frequencies. Finally, we must assume that there are no instances of migration or genetic mutation, so that natural selection is the only force at work.

3.1 Modeling Natural Selection

To simulate a population in which different individuals are more or less likely to reproduce based on their genotypes, we must first define a few terms. Let

- $g =$ number of generations for which the simulation is run,
- $p =$ allele frequency of allele A (0 to 1),
- $q =$ allele frequency of allele B (0 to 1),
- $w_{AA} =$ frequency with which zygotes with genotype AA survive to reproduce,
- $w_{AB} =$ frequency with which zygotes with genotype AB survive to reproduce,
- $w_{BB} =$ frequency with which zygotes with genotype BB survive to reproduce,
- $\overline{w} =$ average fitness for the whole population during a given generation.

Note that for initial allele frequencies of p and q in the gene pool, we expect the resultant zygotes, or offspring, to have genotypes AA, AB, and BB with frequency p^2, $2pq$, and q^2, respectively. For simplicity in modeling, we assume that all individuals which survive to reproduce bear the same number of offspring. Combining the frequency of each genotype with its respective fitness yields the average fitness for the whole population:

$$\overline{w} = p^2 w_{AA} + 2pq w_{AB} + q^2 w_{BB}.$$

The genotypic frequencies of the individuals who survive to reproductive age are then given by the frequency of individuals born with the specified genotype, multiplied by that genotype's associated fitness, divided by the average fitness of the population (to ensure that the new frequencies sum to 1). These frequencies are as follows:

AA	AB	BB
$\frac{p^2 w_{AA}}{\overline{w}}$	$\frac{2pq w_{AB}}{\overline{w}}$	$\frac{q^2 w_{BB}}{\overline{w}}$

Because all of the alleles AA individuals contribute to the gene pool are A and half the alleles AB individuals contribute to the gene pool are A, the new frequency of allele A is

$$p' = \frac{p^2 w_{AA} + pq w_{AB}}{\overline{w}}. \tag{1}$$

Thus, we simply use Equation 1 to calculate the allele frequency in each subsequent generation given knowledge of the previous generation's allele frequency and the fitness of each genotype.

3.2 Natural Selection Terminology

Before investigating the trends in natural selection, we must first establish some terminology.

3.2.1 Favorability

An allele which consistently causes its possessors to have a relatively high fitness is termed *favorable*, and one which consistently causes its possessors to have a relatively low fitness is termed *deleterious*.

3.2.2 Dominance

Five different types of allele relationships are possible: full dominance, incomplete dominance, recessivity, overdominance, and underdominance. If allele A is *dominant*, its presence in a genotype "overrides" any other alleles that might be present; the fitness of genotype AA will be equivalent to the fitness of genotype AB because both genotypes contain an A allele. In this case, allele B is the *recessive* allele, since its presence has less of an effect on fitness than does the dominant allele A. Finally, an allele A is *incompletely dominant* if AB individuals have a fitness somewhere between that of the homozygotes.

Overdominance and underdominance occur less frequently than the relationships described above. *Overdominance* is a phenomenon in which the heterozygote AB has a higher fitness than either of the homozygous genotypes. *Underdominance* occurs when the heterozygote has a lower fitness than either of the homozygous genotypes.

3.3 Investigating Trends in Natural Selection

We first explore the effects of dominance, incomplete dominance, and recessivity on a favorable allele A. Figure 8 illustrates natural selection on a population where the fitness of AA individuals is 1.0, the fitness of BB individuals is 0.8, and the fitness of heterozygous AB individuals varies to make A either dominant, incompletely dominant, or recessive for three different initial frequencies. While we show only the effects on a favorable allele, the inverse of these effects will occur for a deleterious allele.

Figure 8 shows that for both high and intermediate initial frequencies, the recessive favorable allele rises to fixation faster than an incompletely dominant one, which in turn

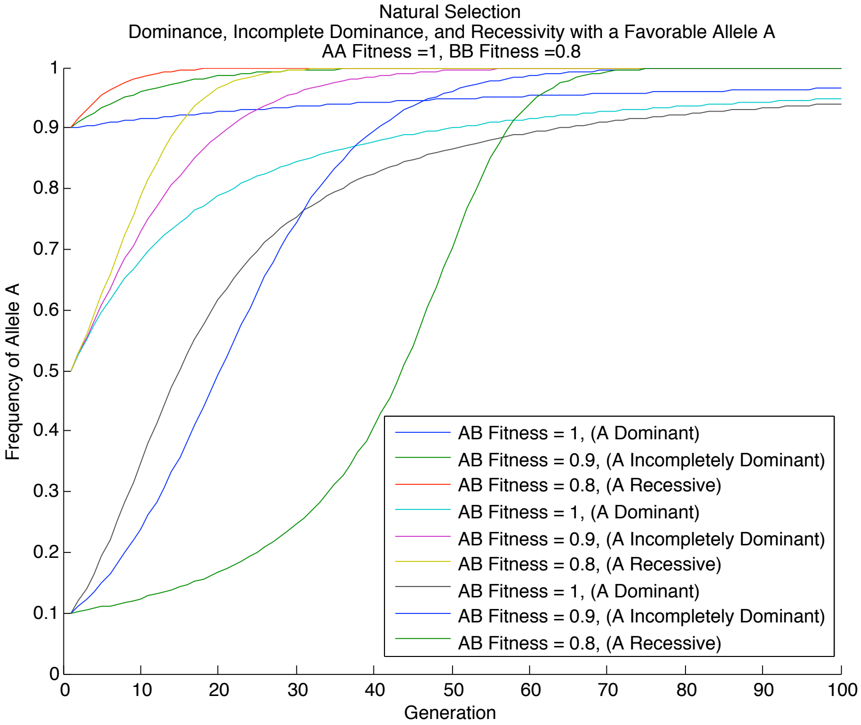


Figure 8: *The effects of natural selection on a population where fitness of homozygous A individuals is 1.0, the fitness of homozygous B individuals is 0.8, and the fitness of heterozygous individuals varies to make a favorable allele A either dominant, incompletely dominant, or recessive for three different initial allele frequencies.*

rises faster than a fully dominant favorable allele. For low initial allele frequencies, however, we see notably different behavior. The frequency of the dominant form of the favorable allele initially rises most quickly, and the frequency of the recessive form initially rises much more slowly, with the incompletely dominant form exhibiting an intermediate initial slope. Strangely, the frequency of the incompletely dominant form of the allele passes that of the dominant form after approximately 25 generations, and the frequency of the recessive form passes it after approximately 50 generations.

Thus, while it initially appears as though the dominant, incompletely dominant, and recessive forms will reach fixation first, second, and third, the trend changes after a number of generations such that they reach fixation last, first, and second. In fact, the iteration with the dominant favorable allele appears as though it never reaches fixation; after 10,000 generations, the dominant allele has only reached a frequency of 0.995. Although counterintuitive, this trend makes sense because the high fitness of the heterozygote (here, $w_{AB} = 1.0$) means that the B allele is maintained within the population to a certain degree. As a result, the frequency of allele A quickly rises when A is favorable and dominant, but the high fitness of the heterozygote eventually slows the upward climb and means that any progress toward fixation of allele A is incredibly slow. This trend occurs for all initial frequencies of A; although it appears on the graph as though A will eventually reach fixation when it is dominant, analysis of the exact frequency of A shows that it never quite reaches a frequency of

1.0, even after 15,000 generations. An analogous situation occurs for a deleterious, recessive allele going toward deletion.

Next we consider the effects of overdominance in a population. Figures 9 and 10 illustrate that overdominance causes the allele frequencies to reach a stable equilibrium frequency which is dependent upon the relative fitnesses of the two homozygous genotypes. When the homozygous fitnesses are equal, as in Figure 9, the equilibrium frequency is 0.5. When the homozygous fitnesses are unequal, as illustrated in Figure 10, the equilibrium frequency is shifted toward the more favorable allele. Thus, we see that overdominance is one case in which two alleles can exist in stable equilibrium within a population.

Finally, we explore the effects of underdominance in a population. Because underdominance is intuitively the inverse of overdominance, it makes sense that underdominance creates an unstable equilibrium frequency. Figures 11 and 12 illustrate cases in which the homozygous fitnesses are equal and unequal, respectively. As with overdominance, the equilibrium frequency depends on the relative fitnesses of the two homozygous genotypes. Initial frequencies that are above or below the equilibrium frequency lead the allele to fixation or deletion, respectively, with initial frequencies equal to the equilibrium frequency remaining there for the duration. Underdominance is another case in which two alleles can exist in equilibrium—albeit an unstable one—within a population.

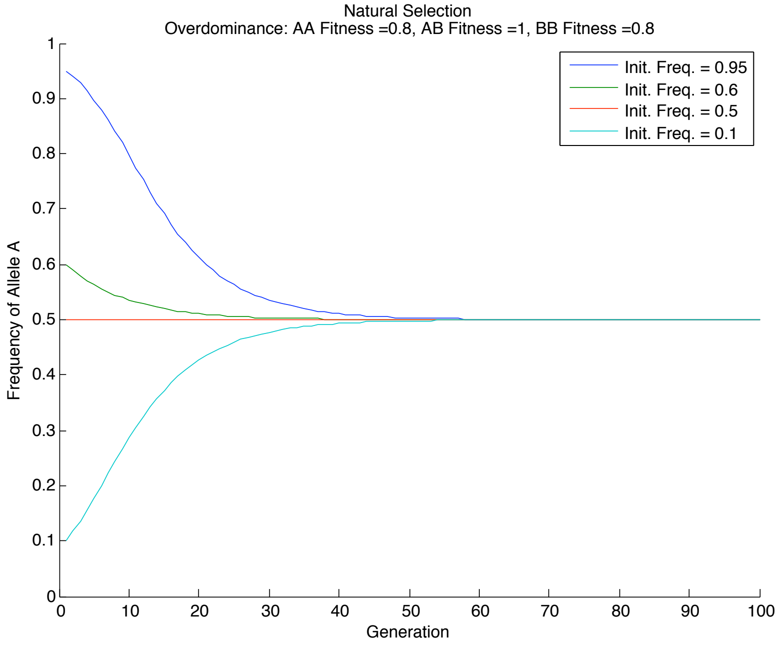


Figure 9: *The effects of natural selection on a population exhibiting overdominance with equivalent fitnesses for the homozygous genotypes. The allele frequencies converge to a stable equilibrium frequency of 0.5.*

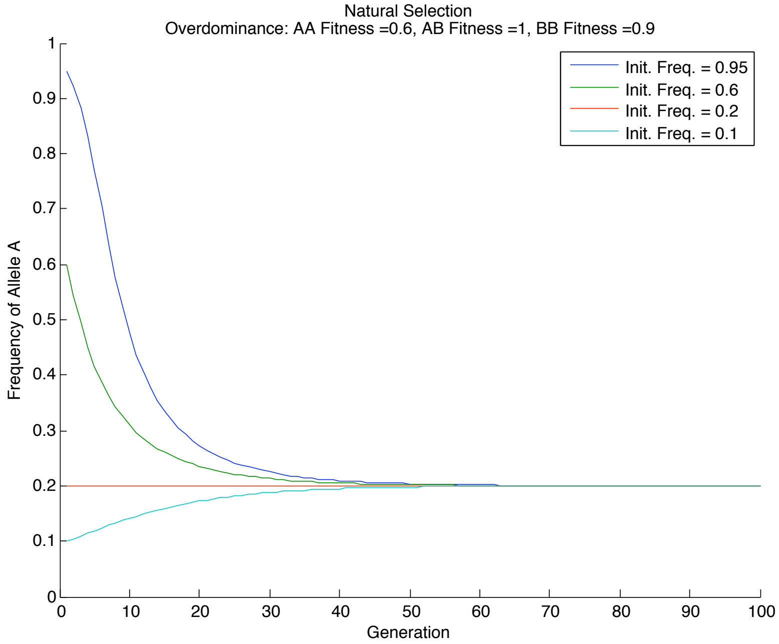


Figure 10: *The effects of natural selection on a population exhibiting overdominance with differing fitnesses for the homozygous genotypes. The allele frequencies converge to a stable equilibrium frequency dependent on the relative fitnesses of the homozygotes.*

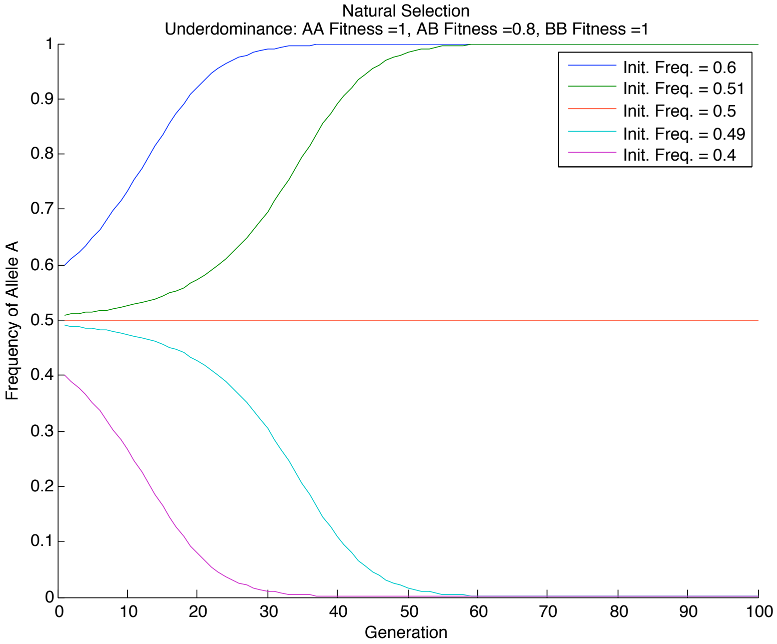


Figure 11: *The effects of natural selection on a population exhibiting underdominance with equivalent fitnesses for the homozygous genotypes. The allele frequencies diverge from an unstable equilibrium of 0.5.*

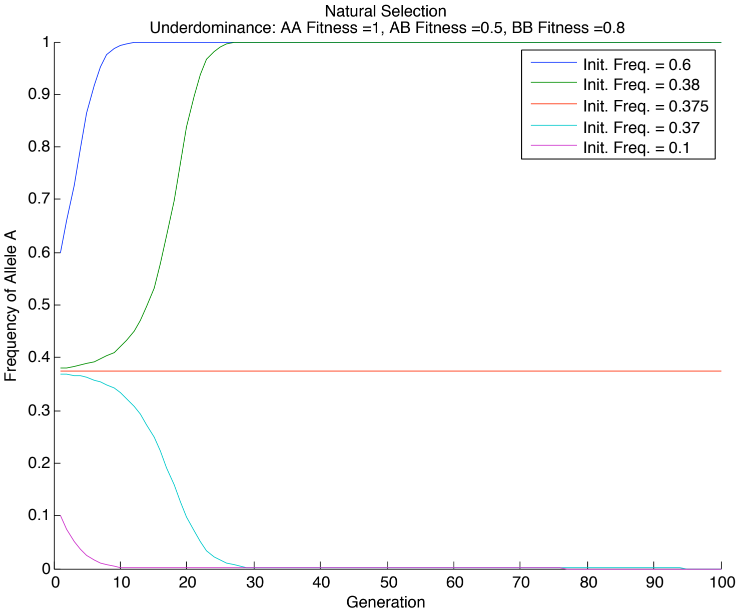


Figure 12: *The effects of natural selection on a population exhibiting underdominance with differing fitnesses for the homozygous genotypes. The allele frequencies diverge from an unstable equilibrium frequency dependent on the relative fitnesses of the homozygotes.*

4 Conclusion

We have now successfully modeled and explored trends in the evolutionary forces of genetic drift and natural selection. Analysis has shown that the two methods used to model genetic drift—the brute force method and the binomial distribution method—are indistinguishable visually and according to their results for the average number of generations until deletion or fixation and for the percentage of deletions and fixations as a function of allele frequency, but that they differ in run time such that the brute force method is faster. Two trends present in genetic drift are that larger population sizes are less affected by drift and that the likelihood of an allele going to fixation via genetic drift is equivalent to its initial frequency in the population.

Natural selection was modeled using a recurrence relationship relating the previous generation's allele frequency to that of the subsequent generation. Analysis of the trends in selection showed that for all initial frequencies of a favorable allele A, the dominant form of the allele does not reach fixation after 15,000 generations. For high and intermediate initial allele frequencies, a favorable allele A reaches fixation first when it is recessive, then incompletely dominant. For low initial frequencies, the order of fixation is incompletely dominant, then recessive. While the reason for the lack of fixation of the favorable dominant allele is understood, it would be interesting to further investigate the reason for the difference in order of fixation of the other two types of alleles at differing initial frequencies. We also found that overdominance creates a stable allele frequency equilibrium, while underdominance creates an unstable one.

In the future, it would be beneficial to perform a rigorous statistical analysis on the performance of the two genetic drift methods in order to determine if they are statistically distinguishable. It would also be interesting to find out what is causing the discrepancy in run times. To do this, we would need to dissect MATLAB's `binornd` function and find the differences in implementation, then run a profiler to identify the most time consuming steps. Because genetic drift and natural selection are not the only evolutionary forces at work on allele frequencies within a population (some others are mutation, migration, and gene linkage), future work could also include modeling some of the other forces and compiling the results together to model a population in which multiple forces are at work. To begin the process, we have generated the graph in Figure 13, which illustrates the effect when both natural selection and genetic drift act upon a population of 1,000 individuals.

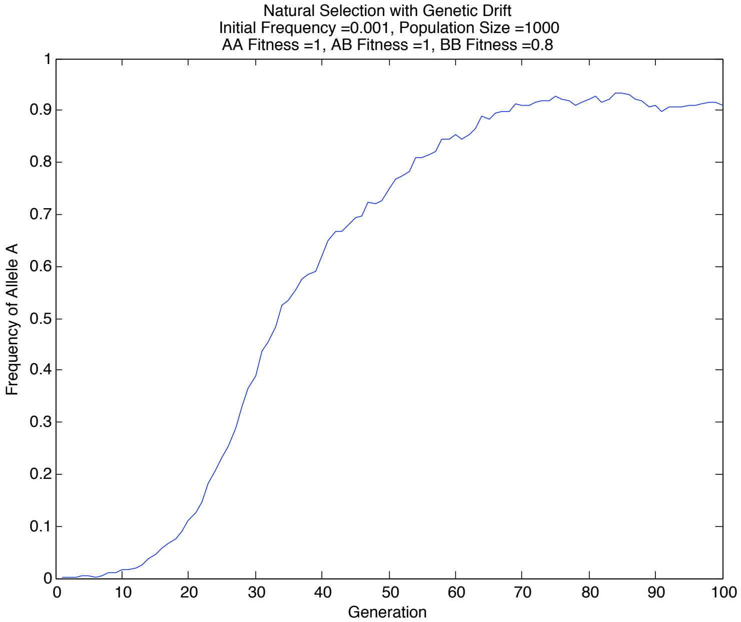


Figure 13: *The effects of both genetic drift and natural selection on a population of 1,000 individuals where allele A is dominant and favorable and the initial allele frequency is 0.001.*

5 Acknowledgements

I would like to thank Professor Darryl Yong for his help developing the idea for this project and formulating the binomial distribution method of modeling genetic drift. I also thank Professor Catherine McFadden, who introduced me to evolutionary biology, for her advice regarding the modeling process and for the material contained in her lectures. Finally, I thank Steven Sloss for his willingness to help me work through the intricacies of MATLAB.