The creation-mutation-selection model: the model and mathematical analysis

Gordon Irlam
gordoni@gordoni.com

Los Altos, California, United States

December 27, 2023

[***Pre-publication draft – feedback solicited***]

Abstract

Despite adaptive mutations being the basis of evolution, many population genetics models of mutation-selection balance have focused on deleterious and neutral mutations. Here, a population genetics model of mutation and selection is described and analyzed that incorporates the creation of new adaptive mutational opportunities. For sexual populations, mean log fitness fitness is determined as a function of the model parameters. For sexual populations and finite genomes a formula for the optimal mutation rate is derived. The optimal mutation rate is found to be one for which the fitness losses associated with positive and negative selection are equal. For asexual populations mean log fitness is found to be proportional to its variance less a constant. This variance is made small by the negative skewness of the log fitness distribution. The sexual model suggests adaptive mutations in sexual populations can't be ignored simply because they are rare. The asexual model suggests asexuality can sometimes incur a large fitness cost.

Keywords: mutation-selection balance, Wright-Fisher model, adaptive mutation, positive selection, deleterious mutation, purifying selection, negative selection, optimal mutation rate.

Introduction

It has been observed that despite adaptive mutations being the basis of evolution, many population genetics models focus on deleterious and neutral mutations, and ignore the effects of positive selection[1, 2].

Possible explanations of the advantage of sex include:

- It allows evolution to proceed more rapidly by enabling the combination of different advantageous and deleterious mutations, the Fisher-Muller hypothesis[3].
- It prevents the build up of deleterious mutations, Muller's ratchet[4].
- It provides the ability to rapidly adapt to reoccurring environmental conditions, including resistance to co-evolving parasites, the Red Queen hypothesis[5].

By incorporating both positive and negative selection, this paper provides a mathematical framework for the investigation of the first two of these explanations, or it can be viewed as standing on its own.

Considerable prior work exists that explores the population genetics of mutation-selection balance. Schiffels et al. enumerated some of the prior work for asexual organisms[6]. A differentiating point for the present work is the use of multiple different mutation related rates: for positive selection a rate at which the environment creates new beneficial mutational opportunities, for negative selection a per organism rate of creation of deleterious mutational opportunities, and a single rate of mutation at these two classes of sites. Fitness is determined by the interplay of the different mutation related rates.

The basic creation-mutation-selection model is presented and mathematically analyzed. In the sexual case fitness is determined by the population size, distribution of selection coefficients, and mutation related rates. An important result for the sexual case is that for a finite genome the optimal mutation rate is one for which fitness losses associated with positive and negative selection are equal. In the asexual case fitness is determined as a function of the selection coefficient, mutation related rates, and the variance, skewness, or higher cumulant of the fitness distribution.

Before commencing it is worth having a clear understanding of what we mean by "fitness", denoted w. Fitness is traditionally defined as the growth rate of a genotype in a population. When organisms are haploid, and the genotype in question is that of the haploid organism, fitness takes on a particularly intuitive form. Fitness is simply the expected relative number of immediate offspring of a haploid organism in some environment. It is an expected value because random stochastic sampling effects may cause the actual observed number to vary from the expected value, but were it possible to repeat the same life cycle in the same environment fitness would equal the mean number of offspring observed. It is relative. An organism with twice the fitness value as some other organism would be expected to have twice as many offspring. Most of the time it is this relative nature of fitness that concerns us, but so we can speak meaningfully of fitness values alone we need to anchor the fitness scale. We choose to discuss fitness here relative to a hypothetical organism that is perfectly adapted to its environment to which we assign a fitness value of 1. This means log

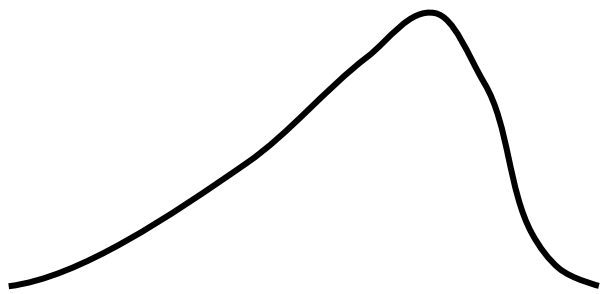

Figure 1: An example of a distribution with negative skewness.

fitness, x, is negative, or zero in the one special case. We do not need to define fitness for diploid organisms, but if we did, it would be as the product of haploid fitness values.

It is also worth reviewing skewness. Skewness is a measure of the asymmetry of a distribution. As shown in Figure 1, a negatively skewed distribution has a long tail to the left of the mean, and a shorter tail to the right. The skewness of a random variable X with mean μ and standard deviation σ is denoted $\tilde{\mu}_3$, and defined as,

$$\tilde{\mu}_3 = \mathrm{E}\left[\left(\frac{X-\mu}{\sigma}\right)^3\right]$$

Non-standardized skewness is termed the third central moment, denoted μ_3, and defined by,

$$\mu_3 = \mathrm{E}\left[(X-\mu)^3\right]$$

Results

Beneficial mutation and selection model

The goal of the model is to strip the organism life cycle down to the bare minimum essential features so that meaningful analysis is possible. The basic model is a Wright-Fisher-like discrete time model consisting of repeating non-overlapping rounds of site creation, mutation, selection, and possibly sexual recombination, in that order. Throughout the process the population size is fixed at N. A gametophyte-centric perspective is taken, meaning that selection is seen as being performed on haploid genomes. Performing selection on haploid genomes is for ease of analysis, but it is not a limitation. Mathematically, it is equivalent to selection occurring on diploid genomes with a heterozygous effect, h, of $\frac{1}{2}$, and a selection coefficient, s_z, for the homozygous state that is 2 times the haploid value. Or more technically for multiplicative fitness, h is $\frac{1}{2-s}$ and s_z is $2s - s^2$, but s is typically small so the distinction is usually immaterial. A monoicous perspective is also taken. Monoicous refers to haploids in which a single organism produces gametes of both sexes[7]. Contrast this with dioicous which refers to haploids that produce sperm and oocytes from

separate male and female gametophytes. And not to be confused with the corresponding terms for the diploid form, dioecous and monoecious. There is no epistasis; multiple selection coefficients in a single organism are assumed to combine multiplicatively as 1 plus their values. Were they to combine additively, frequent use of the term "log fitness", would be replaced by just "fitness". Also note that multiplicative fitness over time in a discrete time model corresponds to additive log fitness over time in a continuous time model.

It may be useful to think of an organism as being a primitive haploid cell.

Site creation. Each organism is assumed to have a set of sites at which beneficial mutations may occur in the future. Site creation is the process of creating new such mutational opportunity sites. Multiple different sites can, and usually will be present in any one organism at any one time. Newly created sites are always created in an unsatisfied state, meaning they exert a negative effect on organismal fitness. Site creation comes in two forms corresponding to positive and negative selection.

Positive site creation. The population is assumed to be embedded in a changing environment. Each time the environment changes matching sites are created in all organisms making up the population, and the fitness of each organism relative to the fitness of the ideal organism in the environment is reduced by a factor of $\frac{1}{1+s_p}$, but at the same time this opens up the possibility of future individual organisms beneficially mutating to bring their fitness up by the amount by which it notionally declined. For all organisms in a particular generation, the number of new sites created in a single generational time-step, Γ, follows a Poisson distribution with mean rate Γ_p, $\Gamma \sim \text{Pois}\left(\Gamma_p\right)$. The associated multiplicative reduction in fitness is,

$$\prod_{i=1,\ldots\Gamma} \frac{1}{1+s_{p,i}}$$

where $s_{p,i}$ is a random variable representing the selection coefficient of the ith new site. On account of selection this may differ from the distribution of existing sites in the population. The environment is assumed to always create new sites, never to delete existing sites, and never to revisit prior sites. This is a simplification of the model. In the real world the environment will likely change quickly at some times and slowly at others. This is not an impediment. What matters is the long term rate of site creation. To make things more physical, site creation might correspond to an environmental change that makes a particular base pair change to the genome now advantageous, with this base pair representing the newly created site.

Negative site creation. Not all mutations are beneficial. Sometimes mutations occur in an individual organism that are deleterious, and will normally be removed from the population either by selection or by chance back mutation. Deleterious mutations are modeled as mutational opportunity sites occurring in each organism according to a Poisson distribution with mean rate of Γ_n per organism. The resulting sites created in each organism are unique to that organism. The selection coefficient associated with the ith new deleterious site in an organism is $s_{n,i}$. Log fitness of the organism is reduced by $\log\left(1+s_{n,i}\right)$, and this is the amount that will be gained should the site be cleared by selection or mutate. In this manuscript selection coefficients for both positive and negative selection are represented by non-negative values, $s_{p,i} \geq 0$ and $s_{n,i} \geq 0$.

Mutation is modeled as occurring at an average rate of μ_{ss} per site per organism per sexual gener-

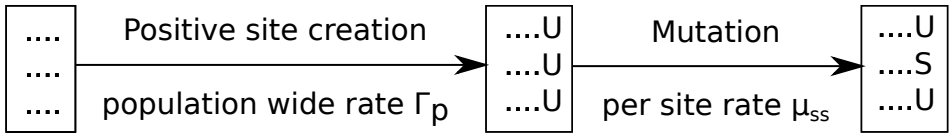

Figure 2: Positive site creation and mutation model. The same positive site gets created in all 3 organism genomes, whereas mutation occurs in each site independently. "U" represents an unsatisfied site, "S" a satisfied site, and "." is an arbitrary site.

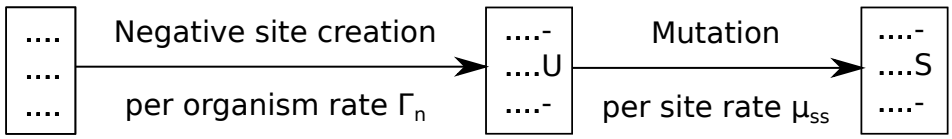

Figure 3: Negative site creation and mutation model. The creation of negative sites occurs in each of the 3 organism genomes independently. Mutation also occurs in each site independently. "U" represents an unsatisfied site, "S" a satisfied site, '-' is a spacer, and "." is an arbitrary site or spacer. The spacer is to preserve alignment of sites, which is only significant for recombination in the sexual model.

ation. New sites may be created, but once a site has been satisfied by a mutation the satisfied site will no longer exert a negative effect on fitness, nor will it ever become unsatisfied.

Neutral mutations are considered as a special subset of the sites created by positive and negative site creation, for which $s < \frac{1}{N}$, and random genetic drift dominates over selection. The neutral site creation rate, Γ_0, is defined as,

$$\Gamma_0 = \int_0^{\frac{1}{N}} \Gamma_p(s) + \Gamma_n(s) ds$$

where $\Gamma_p(s)$ and $\Gamma_n(s)$ are site creation rate densities associated with positive and negative site creation.

Sometimes we will speak of true positive sites. These are positive sites excluding neutral positive sites. The true positive site creation rate, Γ_p^*, is,

$$\Gamma_p^* = \int_{\frac{1}{N}}^{\infty} \Gamma_p(s) ds$$

Similarly, the true negative site creation rate Γ_n^*, is,

$$\Gamma_n^* = \int_{\frac{1}{N}}^{\infty} \Gamma_n(s) ds$$

The positive and negative site creation and mutation models are illustrated in Figures 2 and 3 respectively.

Selection is modeled by drawing a new population of size N with replacement where each existing organism is selected from with relative probability w, the fitness of that organism. Note, we take a gametophyte-centric perspective, not simply a gamete-centric perspective, meaning a single haploid genome can give rise to multiple offspring.

Sexual recombination. Sex, if present, is modeled as drawing a new population of size N with replacement from the equiprobable universe of all possible crosses of the population, including self crosses. The resulting organism shares all the sites that are present in both parents, and has a 50% chance of sharing each site present in just one parent. This gets to the heart of what sex involves, the recombination of genomes. Splitting the haploid population into separate mating types and prohibiting self crosses to match dioicous species would complicate the analysis, and would be unlikely to significantly change the results. It would do little to slow the rapid rate at which sex brings genomes to an equilibrium. Were we a monoicous species, such as many bryophytes, this entire life cycle would appear natural. It is only the inability of the haploid form to undergo mitosis in humans and the long lived nature of the diploid form that make this mode of sexuality seem at all unusual. The diploid organism appears in this analysis but briefly, and only during sex. As previously mentioned this isn't a limitation, assuming $h = \frac{1}{2}$.

The goal for both the sexual and asexual cases is to determine the population's fitness, and along the way hopefully understand the factors contributing to it.

An overarching principle is that an organism or a population with many unsatisfied sites will experience a high rate of site satisfaction, rapidly reducing the number of sites. While an organism or population with few sites will experience a lower rate of site satisfaction slowly reducing the number of sites. Thus there exists an equilibrium number of sites for which the rate of creation of new sites by the environment is balanced by the satisfaction of existing sites.

Analysis of the sexual model

Despite containing an additional step, the sexual model is easier to analyze than the asexual model. The reason for this is it is assumed that sex will maintain linkage equilibrium between the different unsatisfied sites. In the real world, there is linkage between nearby locations on a chromosome. As a simplification of the model, this linkage is ignored. Initially, to simplify the analysis, it will be assumed that all sites have the same selection coefficient, s.

Positive selection for a non-extremely large or small population

For an asexual population the probability of a mutation fixing is approximately $1 - e^{-2s}$[8]. This is for successive rounds of selection with selection coefficient of s. For the sexual population we have rounds comprising selection with selection coefficient s immediately followed by selection with selection coefficient 0. This later selective round is the result of choosing parents and sites from the previously selected population. Thus the average selective force is approximately $\frac{s}{2}$, and the probability of fixation for the sexual population is $1 - e^{-s}$.

Provided that we don't have multiple mutations of a single site racing to fixation concurrently, we can use basic queuing theory to determine mean log fitness.

For the sexual model, the probability of a mutation to an adaptive site fixing is approximately $1 - e^{-s}$, and the number of mutations of a single site per generation is $N\mu_{ss}$. An organism with k unsatisfied sites has log fitness $x = -k\log(1+s)$. Consequently if there are an average of $\overline{k}$ sites per organism,

$$\Gamma_p \approx \overline{k}(1 - e^{-s})N\mu_{ss}$$

Hence,

$$\begin{aligned}\overline{x} &\approx -\frac{\Gamma_p \log(1+s)}{N\mu_{ss}(1-e^{-s})} \\ &\approx -\frac{\Gamma_p(s - \frac{s^2}{2} + \cdots)}{N\mu_{ss}(s - \frac{s^2}{2} + \cdots)} \\ &\approx -\frac{\Gamma_p}{N\mu_{ss}} \qquad (1)\end{aligned}$$

Equation 1 applies when the fixation time is much smaller than the expected time between consecutive fixation attempts for a single site. Provided $N \gg \frac{1}{s}$, the fixation time is $\frac{\log Ns}{s}$, as determined by applying the analysis of Desai and Fisher to a sexual population[9]. And the probability of fixation is s, making the time between consecutive fixation attempts is $\frac{1}{Ns\mu_{ss}}$. So equation 1 applies when,

$$\frac{1}{s} \ll N \text{ and } N\log Ns \ll \frac{1}{\mu_{ss}} \qquad (2)$$

Positive selection for an extremely large population

Suppose the population is large enough for a group of matching sites to have multiple satisfying mutations occur to it in a single time-step,

$$N \gg \frac{1}{\mu_{ss}}$$

Then we can reasonably speak in terms of the probability of the site still existing in a randomly chosen organism after some time period.

Define the reduction in fitness as a result of a site going unsatisfied by $s_{rev} = 1 - \frac{1}{1+s}$. For a single generation, the probability of a single site going unsatisfied by a mutation, p_u, is $e^{-\mu_{ss}}$. Normally μ_{ss} is very small, so $p_u \approx 1 - \mu_{ss}$.

Let $a_{s,i}$ be the probability that a single specific unsatisfied positive site created i generations ago is present in any one particular organism now. When $a_{s,i-1}$ is close to 1, all of the other organisms will have the same site, s thus won't have any relative effect, and $a_{s,i}$ will be given by $a_{s,i-1}$ times

the probability of the site surviving, or $p_u a_{s,i-1}$. When $a_{s,i-1}$ is close to 0, s will decrease the probability of the site surviving, and $a_{s,i}$ will be given by $p_u(1 - s_{rev})a_{s,i-1}$. More generally, we will interpolate between these two values based on $a_{s,i-1}$,

$$\begin{aligned} a_{s,i} &= a_{s,i-1} p_u a_{s,i-1} + (1 - a_{s,i-1})\, p_u(1 - s_{rev}) a_{s,i-1} \\ &= p_u a_{s,i-1} \left(1 - s_{rev} + s_{rev} a_{s,i-1}\right) \end{aligned}$$

and we also have $a_{s,0} = 1$. Thus $a_{s,i}$ is given by a simple recurrence relation. Despite the simplicity of this recurrence relation, it appears to have no known solution. This isn't an impediment however, as it is possible to use a computer to rapidly compute the $a_{s,i}$ values.

Recalling that an average of Γ_p new sites are created per generation, the mean total number of sites present in any one organism, $\overline{k}$, is given by,

$$\overline{k} = \Gamma_p \sum_{i=1,\ldots\infty} a_{s,i}$$

So that,

$$\begin{aligned} \overline{x} &= -\overline{k} \log(1 + s) \\ &= -\Gamma_p \log(1 + s) \sum_{i=1,\ldots\infty} a_{s,i} \end{aligned} \tag{3}$$

Since each site is independent, and the chance of a given site being present in a particular organism follows a Bernoulli, or $\mathrm{Binom}\,(1, a_{s,i})$ distribution, with variance $a_{s,i}\,(1 - a_{s,i})$, the variance in the number of sites, σ_k^2, is given by,

$$\sigma_k^2 = \Gamma_p \sum_{i=1,\ldots\infty} a_{s,i}\,(1 - a_{s,i}) \tag{4}$$

Negative selection

Negative selection is easier to analyze than positive selection.

Let $b_{s,i}$ be the probability that a single specific negative site created i generations ago is present in any one particular organism now. Initially we have $b_{s,0} = \frac{1}{N}$. The chance of the site surviving one generation is $p_u\,(1 - s_{rev})$, so,

$$
\begin{aligned}
b_{s,i} &= p_u \left(1 - s_{rev}\right) b_{s,i-1} \\
&= \frac{\left(\frac{p_u}{1+s}\right)^i}{N}
\end{aligned}
$$

The mean number of negative sites created each generation is $N\Gamma_n$, so that,

$$
\begin{aligned}
\overline{k} &= N\Gamma_n \sum_{i=1,\ldots\infty} b_{s,i} \\
&= \Gamma_n \frac{\frac{p_u}{1+s}}{1 - \frac{p_u}{1+s}} \\
&\approx \frac{\Gamma_n}{s} \text{ provided } \mu_{ss} \ll s
\end{aligned}
$$

As a result,

$$
\begin{aligned}
\overline{x} &= -\overline{k} \log(1+s) \\
&\approx -\frac{\Gamma_n \log(1+s)}{s} \\
&\approx -\Gamma_n \text{ provided } s \text{ small}
\end{aligned}
\tag{5}
$$

Each site is independent, and follows a Bernoulli distribution, thus the variance in the number of sites is,

$$
\begin{aligned}
\sigma_k^2 &= N\Gamma_n \sum_{i=1,\ldots\infty} b_{s,i} \left(1 - b_{s,i}\right) \\
&\approx \frac{\Gamma_n}{s} \text{ provided } N \text{ large and } \mu_{ss} \ll s
\end{aligned}
$$

That is, the variance in the number of negative sites across all organisms is approximately equal to the mean number of negative sites across all organisms.

Positive and negative selection for a tiny population or a tiny selection coefficient

The analysis we have performed so far breaks down if random drift dominates over selection. That is we have implicitly assumed that $N \gg \frac{1}{s}$. If $N \ll \frac{1}{s}$, the sites are neutral sites[10]. In this case, the effect of selection is negligible, and we can again use basic queuing theory to analyze the situation. The rate of arrival of sites must equal the rate at which they are satisfied by mutation,

$$N\Gamma_0 = N\overline{k}\mu_{ss}$$

So that,

$$\begin{aligned}\overline{x} &= -\frac{\Gamma_0 \log(1+s)}{\mu_{ss}} \\ &\approx -\frac{\Gamma_0 s}{\mu_{ss}}\end{aligned} \tag{6}$$

Consequently, depending on the site creation rate, the fitness loss associated with neutral sites can be substantial. They are neutral with respect to selection, not fitness.

Despite the potentially high cost of neutral sites, we may be able to ignore them when comparing the fitness of asexual and sexual populations. This is because for a single s value, the queuing theory analysis performed here applies equally to a sexual or asexual population. This is a consequence of drift dominating over selection.

Multiple selection coefficients

Sexual recombination breaks up any linkage between sites, which makes the effects of different selection coefficients independent. As a result the sexual theory that has been developed should apply equally to a single selection coefficient or a range of selection coefficients. It also means it should be possible to combine the effects of positive and negative selection.

For a non-extremely large population, replacing Γ_p and Γ_n by their densities $\Gamma_p(s)$ and $\Gamma_n(s)$, gives the rough general formula,

$$\overline{x} \approx -\int_0^{\frac{1}{N}} \frac{\Gamma_p(s)+\Gamma_n(s)}{\mu_{ss}} ds - \int_{\frac{1}{N}}^{\infty} \frac{\Gamma_p(s)}{N\mu_{ss}} + \frac{\Gamma_n(s)\log(1+s)}{s} ds$$

Where a more accurate formula would show some blending of the integrands in the region $s = \frac{1}{N}$.

This equation shows the increase in the effect of neutral selection coefficients: by a factor of N for positive selection, and by a factor of $\frac{1}{\mu_{ss}}$ for negative selection. Also, there is a modest reduction in the effect of large selection coefficients for negative selection.

We could also write this equation as,

$$\overline{x} \approx -\frac{\Gamma_0}{\mu_{ss}} - \frac{\Gamma_p^*}{N\mu_{ss}} - \Gamma_n^* \tag{7}$$

In which we are ignoring the modestly reduced impact for negative selection with large s values.

It is interesting to note that fitness is largely independent of the distribution of fitness effects for new sites, depending only on aggregate rates of new site creation. This is presumably because satisfying mutations have a fixation probability of approximately s, meaning that smaller s values will take longer before they are fixed, and thus the sites exerts their fitness reducing effects for a longer period of time.

Properties of positive selection for a very large population and negative selection

Since consecutive $a_{s,i}$ are almost identical, loosely speaking, the central limit theorem will apply, and the number of sites present in a specific organism should approximately follow a normal distribution. The same is also true for $b_{s,i}$ provided that s is small. Fitness is given by $\frac{1}{1+s}$ raised to the power of k. Consequently fitness is expected to approximately follow a log-normal distribution.

Two other attributes worth considering are the number of matching pairs of sites and total differing sites that occur in the diploid form when two organisms cross, m and d. Here we will just consider the case of positive selection for a very large population, but the results for negative selection follow a fairly similar pattern.

$$\mathrm{E}\left[m\right] = \Gamma_p \sum_{i=1,\ldots\infty} a_{s,i}^2$$

Noting that differing sites could occur in either one of the parent organisms,

$$\begin{aligned} \mathrm{E}\left[d\right] &= 2\Gamma_p \sum_{i=1,\ldots\infty} a_{s,i}\left(1 - a_{s,i}\right) \\ &= 2\sigma_k^2 \end{aligned} \tag{8}$$

Convergence of variance for a sexual population

Distribution of a distribution lemma: Let Dist_1 be a non-negative distribution with mean μ_1. Let $\mathrm{Dist}_2(d)$ be a distribution parameterized by d with mean 0 and variance ad for some non-negative constant a. Then sampling Dist_1, yielding a value d, followed by sampling the distribution $\mathrm{Dist}_2(d)$ yielding a value x, creates a new distribution with,

$$\mathrm{Var}\left[x\right] = a\mu_1$$

Proof:

$$
\begin{aligned}
\underset{d\in\mathrm{Dist}_1, x\in\mathrm{Dist}_2(d)}{\mathrm{Var}}[x] &= \underset{d\in\mathrm{Dist}_1, x\in\mathrm{Dist}_2(d)}{\mathrm{E}}\left[x^2\right] - \underset{d\in\mathrm{Dist}_1, x\in\mathrm{Dist}_2(d)}{\mathrm{E}}[x]^2 \\
&= \underset{d\in\mathrm{Dist}_1}{\mathrm{E}}\left[\underset{x\in\mathrm{Dist}_2(d)}{\mathrm{Var}}[x] + \underset{x\in\mathrm{Dist}_2(d)}{\mathrm{E}}[x]^2\right] - \underset{d\in\mathrm{Dist}_1}{\mathrm{E}}[0]^2 \\
&= \underset{d\in\mathrm{Dist}_1}{\mathrm{E}}[ad] \\
&= a\mu_1
\end{aligned}
$$

Let m be the number of matching sites, and d be the total number of differing sites when two organisms cross.

For an arbitrary state, not necessarily the steady state equilibrium, and for an infinite or finite population, the distribution of the number of sites in the child is given by $m + \mathrm{Binom}(d, 0.5)$. If the total number of sites in the parents are k_{p1} and k_{p2}, and the number of sites in the child is k_c,

$$
\mathrm{E}[k_c] = \mathrm{E}[m] + \frac{\mathrm{E}[d]}{2}
$$

$$
\begin{aligned}
\mathrm{Var}[k_c] &= \mathrm{Var}\left[\frac{k_{p1} + k_{p2} - d}{2} + \mathrm{Binom}(d, 0.5)\right] \\
&= \frac{\mathrm{Var}[k_{p1}]}{4} + \frac{\mathrm{Var}[k_{p2}]}{4} + \mathrm{Var}\left[\mathrm{Binom}(d, 0.5) - \frac{d}{2}\right] && (9) \\
&= \frac{\sigma_k^2}{2} + \frac{\mathrm{E}[d]}{4} && (10)
\end{aligned}
$$

The lack of covariance terms on line 9 is justified because the sites in k_{p1} and k_{p2} are independent, and the final term is a symmetric binomial distribution centered about the origin so that for any k_{p1} and k_{p2} values it could just as easily be greater than zero as it could be less than zero; in other words the correlation is zero. The last line follows by application of the distribution of a distribution lemma, noting that for fixed value of d, $\mathrm{Var}\left[\mathrm{Binom}(d, 0.5) - \frac{d}{2}\right] = \mathrm{Var}[\mathrm{Binom}(d, 0.5)] = \frac{d}{4}$.

Since $\mathrm{Var}[k_c]$ is the σ_k^2 of the new generation, it follows that log fitness variance in the sexual case will rapidly converge to $\frac{\mathrm{E}[d]}{2}(\log(1+s))^2$. For a very large population this agrees with equation 8.

Finite sites model with four bases

So far the discussion has been very abstract. If we are to apply the model to biological systems we need to make it more realistic. In particular each site has four different possible base pair values, there is a limit on the number of new sites that can exist at any one time due to the finite length of the genome, and that back mutation is possible.

Let μ_{bs} be the DNA mutation rate for the species per base pair per sexual generation. Under the rough assumptions that there is only one correct mutation that satisfies a new positive site with the other two mutations leaving the site on average no better and no-worse,

$$\mu_{ss} = \frac{1}{3}\mu_{bs} \tag{11}$$

Let L be the length of the genome, and l_0 be the fraction of L for which mutations are neutral. Then, $\Gamma_0 = l_0 L \mu_{bs}$. It is tempting to plug this value into equation 7, and derive the log fitness loss associated with neutral sites for sexual populations as $-3l_0L$. However, this ignore the possibilities of sites being lost because of the finite length of the genome and the possibility of back mutation.

Let the mean selection coefficient for neutral base pair sites be s_0. Neutrality means drift dominates over selection, and so at any given time $\frac{3}{4}$ of neutral base pair sites will be unsatisfied. Thus the fitness loss associated with neutral sites will approximately be $\frac{3l_0Ls_0}{4}$.

Let l_n be the fraction of L that is under the control of negative selection. Then,

$$\Gamma_n^* = l_n L \mu_{bs} \tag{12}$$

Thus based on equation 7, for a realistic sexual population,

$$\overline{x} \approx -\frac{3l_0Ls_0}{4} - \frac{3\Gamma_p^*}{N\mu_{bs}} - l_n L \mu_{bs} \tag{13}$$

where the three terms on the right hand side give the fitness losses associated with neutral, true positive, and true negative sites, respectively.

The optimal sexual per site mutation rate

It has been hypothesized that natural selection will drive the sexual mutation rate down to the lowest possible level taking into account the cost of producing such low mutation rates or that a lower bound is set by random genetic drift[11].

Here we instead hypothesize that evolution tunes the mutation rate per generation so that it is as large as possible to maximize the ability to adapt and hence minimize the fitness cost associated with positive selection, but not so large that negative, or purifying, selection incurs a heavy cost.

For a particular value of g, if Γ_p^*, N, l_0L, l_nL, and s_0 are all independent of μ_{bs}, equation 13 takes on its maximal value when,

$$\frac{3\Gamma_p^*}{N\mu_{bs}^2} \approx l_n L$$

$$\mu_{bs} \approx \sqrt{\frac{3\Gamma_p^*}{N l_n L}}$$

$$\mu_{bs} \approx \frac{3\Gamma_p^*}{N\Gamma_n^*} \text{ by equation 12} \tag{14}$$

of,

$$\overline{x} \approx -\frac{3 l_0 L s_0}{4} - \Gamma_n^* - \Gamma_n^*$$

In other words, at the maximum fitness value, the fitness losses coming from positive selection are equal to those coming from negative selection. This is an important result, as often in mutation-selection models positive selection is ignored.

Analysis of the asexual model

Fitness cumulant mutational moment relationship

This subsection reproduces the work of Gerrish and Sniegowski[12, 13], simplifying it by assuming the mutational outflux is total, and that mutational effects depend only on time. For related work see Good and Desai[14]. An infinite population is assumed. Both time, t, and log fitness, x, are assumed to be approximated by continuous variables. $u(x,t)$ is the fitness distribution probability density function of the population. $g(\phi,t)$ is the site creation and mutational log effect distribution probability density function. That is, the probability density of site creation or a mutation occurring that has log effect ϕ, expressed as an instantaneous rate per unit of time. The instantaneous rate of population increase due to selection is x. To maintain the probability distribution this means the rate of change in u at x due to selection must be,

$$(x - \overline{x})\, u(x,t)$$

The rate of influx at x as a result of site creation and mutation is given by,

$$\int_{-\infty}^{\infty} u(x-\phi,t) g(\phi,t) d\phi$$

And the rate of mutational outflux at x in our case is considered to be total,

$$u(x,t)$$

If required, any lack of mutation at x will be represented by use a Dirac delta like function $\delta(\phi)$ at $g(0,t)$, $\delta(\phi) = 0$ for $\phi \neq 0$ and $\int_{-\infty}^{\infty} \delta(\phi) d\phi = 1$.

Combining the effects of selection and site creation and mutation gives,

$$\frac{\partial}{\partial t} u(x,t) = (x - \overline{x})\, u(x,t) + \int_{-\infty}^{\infty} u(x-\phi,t) g(\phi,t) d\phi - u(x,t) \tag{15}$$

An exact solution to equation 15 is not sought because any real world approximation to equation 15 is going to experience a degree of stochasticity on account of a finite population. Instead the equation is recast into a form that makes this stochasticity explicit. Defining,

$$\Pi(\theta,t) = \int_{-\infty}^{\infty} e^{\theta(x-\overline{x})} u(x,t) dx$$

$$\psi(\theta,t) = \log \Pi(\theta,t)$$

leads to the equation,

$$\frac{\partial}{\partial t} \psi(\theta,t) + \theta \frac{\partial}{\partial t} \overline{x} = \frac{\partial}{\partial \theta} \psi(\theta,t) + M(\theta,t) - 1 \tag{16}$$

where

$$M(\theta,t) = \sum_{i=0}^{\infty} \frac{1}{i!} m_{i,\phi}(t) \theta^i$$

is the moment generating function of the mutational effects distribution, and,

$$m_{i,\phi}(t) = \int_{-\infty}^{\infty} \phi^i g(\phi,t) d\phi$$

is the ith moment.

Equation 16 can be verified by observing,

$$\frac{\partial}{\partial t} \psi(\theta,t) = \frac{1}{\Pi(\theta,t)} \int_{-\infty}^{\infty} e^{\theta(x-\overline{x})} \frac{\partial}{\partial t} u(x,t) dx - \theta \frac{\partial}{\partial t} \overline{x}$$

$$\frac{\partial}{\partial \theta} \psi(\theta,t) = \frac{1}{\Pi(\theta,t)} \int_{-\infty}^{\infty} e^{\theta(x-\overline{x})} (x - \overline{x})\, u(x,t) dx$$

$$\frac{1}{\Pi(\theta,t)}\int_{-\infty}^{\infty} e^{\theta(x-\overline{x})}\int_{-\infty}^{\infty} u(x-\phi,t)g(\phi,t)d\phi dx = \int_{-\infty}^{\infty} e^{\theta\phi}g(\phi,t)d\phi$$
$$= \int_{-\infty}^{\infty}\sum_{i=0}^{\infty}\frac{(\theta\phi)^i}{i!}g(\phi,t)d\phi$$
$$= M(\theta,t)$$

It is worth noting in equation 16 that $\frac{\partial}{\partial\theta}\psi(\theta,t)$ represents the effect of selection, while $M(\theta,t)-1$ represents the effect of mutation.

ψ is the central cumulant generating function of u,

$$\psi(\theta,t) = \sum_{i=1}^{\infty}\frac{\theta^i}{i!}\kappa_i$$

where

$$\kappa_1 = 0$$
$$\kappa_2 = \sigma_x^2$$
$$\kappa_3 = \mu_{3,x}$$

Differentiating equation 16 once, twice, and i times ($i > 1$) with respect to θ, and evaluating each time at $\theta = 0$,

$$\frac{\partial}{\partial t}\overline{x} = \kappa_2 + m_{1,\phi} \tag{17}$$
$$\frac{\partial}{\partial t}\kappa_2 = \kappa_3 + m_{2,\phi} \tag{18}$$
$$\frac{\partial}{\partial t}\kappa_i = \kappa_{i+1} + m_{i,\phi}$$

In which the first term on the right hand side represents the effect of selection, and the second term on the right hand side represents the effect of mutation. So in general each log fitness cumulant is driven by the value of the one higher cumulant offset by the overall mutational effect moment.

Dynamic equations

It is assumed all sites have the same selection coefficient, s. This assumption is a limitation of the asexual model brought about by the distribution of fitness effects for an organism depending on the organism's fitness. Let,

$$S = \log(1+s)$$

Consider mutation. An organism with all sites satisfied has a log fitness of zero, so the number of unsatisfied sites is $-\frac{x}{S}$. The rate at which each site might be satisfied is μ_{ss}. And the effect of mutation is to shift log fitness by S, resulting in a $\delta(\phi - S)$ spike in the mutational probability density function. Multiplying these factors together contributes the term $-\frac{\mu_{ss}x}{S}\delta(\phi - S)$ to $g(\phi, t)$.

Negative sites are created at the rate Γ_n, and shift log fitness by $-S$, contrbutinghte term $\Gamma_n\delta(\phi+S)$ to $g(\phi, t)$. So the continuous time equivalent of negative site creation is,

$$g(\phi, t) = -\frac{\mu_{ss}x}{S}\delta(\phi - S) + a\delta(\phi) + \Gamma_n\delta(\phi + S) \text{ for negative selection}$$

where $a = 1 - \Gamma_n + \frac{\mu_{ss}x}{S}$.

For positive selection, things get more complex. This is because the occurrence of each new site is not independent across organisms. Instead positive site creation is modeled as resulting in a continuous reduction in fitness at a rate of $\Gamma_p S$. Let $\Delta_{a,b}$ be a generalization of the Dirac delta function with the following properties: $\Delta_{a,b}(\phi) = 0$ for $\phi \neq 0$, $\int_{-\infty}^{\infty} \Delta_{a,b}(\phi)d\phi = a$, $\int_{-\infty}^{\infty} \phi\Delta_{a,b}(\phi)d\phi = b$, $\int_{-\infty}^{\infty} \phi^i\Delta_{a,b}(\phi)d\phi = 0$ for $i > 1$. It may be thought of as the limiting case of a very cleverly shaped spike. The continuous time equivalent of discrete site creation is for $g(\phi, t)$ to exhibit a $\Delta_{\cdot,-\Gamma_p S}$ shaped spike at $\phi = 0$. It follows that,

$$g(\phi, t) = -\frac{\mu_{ss}x}{S}\delta(\phi - S) + \Delta_{\cdot,-\Gamma_p S}(\phi) \text{ for positive selection}$$

Assume $x/\overline{x} \approx 1$, that is all organisms have broadly similar fitness in comparison to the full range of possible fitness values. It follows that,

$$\begin{aligned} m_{1,\phi} &= \int_{-\infty}^{\infty} \phi g(\phi, t)d\phi \\ &\approx \begin{cases} -\mu_{ss}\overline{x} - \Gamma_p S & \text{for positive selection} \\ -\mu_{ss}\overline{x} - \Gamma_n S & \text{for negative selection} \end{cases} \end{aligned} \tag{19}$$

and,

$$\begin{aligned} m_{2,\phi} &= \int_{-\infty}^{\infty} \phi^2 g(\phi, t)d\phi \\ &\approx \begin{cases} -\mu_{ss}\overline{x}S & \text{for positive selection} \\ -\mu_{ss}\overline{x}S + \Gamma_n S^2 & \text{for negative selection} \end{cases} \end{aligned} \tag{20}$$

and,

$$
\begin{aligned}
m_{i,\phi} &= \int_{-\infty}^{\infty} \phi^i g(\phi, t) d\phi \\
&\approx \begin{cases} -\mu_{ss}\overline{x}S^{i-1} & \text{for positive selection} \\ -\mu_{ss}\overline{x}S^{i-1} + \Gamma_n(-S)^i & \text{for negative selection} \end{cases} \\
&\geq 0 \text{ for positive selection}
\end{aligned}
$$

Combining equations 19 and 20 with equations 17 and 18,

$$
\frac{\partial}{\partial t}\overline{x} \approx \begin{cases} \sigma_x^2 - \mu_{ss}\overline{x} - \Gamma_p S & \text{for positive selection} \\ \sigma_x^2 - \mu_{ss}\overline{x} - \Gamma_n S & \text{for negative selection} \end{cases} \tag{21}
$$

and,

$$
\frac{\partial}{\partial t}\sigma_x^2 \approx \begin{cases} \mu_{3,x} - \mu_{ss}\overline{x}S & \text{for positive selection} \\ \mu_{3,x} - \mu_{ss}\overline{x}S + \Gamma_n S^2 & \text{for negative selection} \end{cases} \tag{22}
$$

Suppose $\mu_{3,x}$ is determined, and $\overline{x}$ is less negative than the equilibrium solution for equation 22. Then by this equation, $\frac{\partial}{\partial t}\sigma_x^2$ will be negative and σ_x^2 will decrease. σ_x^2 will then cause $\overline{x}$ to change. By equation 21 the change would either immediately, or eventually, make $\overline{x}$ more negative. This would continue until $\mu_{3,x}$ and $\overline{x}$ achieved equilibrium. The reverse argument applies when $\overline{x}$ is more negative than its equilibrium solution. Thus non-standardized log fitness skewness causes the change in log fitness variance, and log fitness variance causes the change in mean log fitness.

Steady state solution

At the steady state the partial derivatives with respect to t will be zero, producing from equations 21 and 22,

$$
\overline{x} \approx \begin{cases} \frac{\sigma_x^2 - \Gamma_p S}{\mu_{ss}} & \text{for positive selection} \\ \\ \frac{\sigma_x^2 - \Gamma_n S}{\mu_{ss}} & \text{for negative selection} \end{cases} \tag{23}
$$

and,

$$
\overline{x} \approx \begin{cases} \frac{\mu_{3,x}}{\mu_{ss}S} & \text{for positive selection} \\ \\ \frac{\mu_{3,x} + \Gamma_n S^2}{\mu_{ss}S} & \text{for negative selection} \end{cases} \tag{24}
$$

Equation 23 shows that mean log fitness is proportional to the variance in log fitness less a constant. Both equations 23 and 24 have μ_{ss} in the denominator, which is typically a very small value, potentially leading to a large negative value for $\overline{x}$, and an extremely small value for fitness, w.

Consulting equation 24. since $\overline{x} \leq 0$ it follows that $\tilde{\mu}_{3,x} \leq 0$. This ignores stochastic effects. For relatively small population sizes $\mu_{3,x}$ can be expected to fluctuate substantially over time, sometimes even becoming positive. Consequently the force exerted by $\mu_{3,x}$ on σ_x^2 (equation 22) will sometimes be large, and sometimes be small. What matters though for the long run value of σ_x^2 is the long run mean value of $\mu_{3,x}$, the fluctuations don't matter.

Higher order equations

Non-standardized skewness of the log fitness distribution, κ_3, is driven by $\kappa_4 + m_{3,\phi}$. For positive selection, since $m_{3,\phi} \geq 0$, it follows that at the steady state κ_4 must be negative or zero. That is the effect of selection is to act to make skewness more negative, while mutation acts to make it more positive.

It is certainly possible to differentiate equation 16 further and derive additional relationships. κ_4 is driven by $\kappa_5 + m_{4,\phi}$. And so on. Each cumulant is driven by selection based on the value of a higher cumulant and offset by the force of mutation. Beyond this though, these higher order equations appear to add little additional understanding.

These higher order equations can also only be taken so far. They overlook the discrete time nature of the original model, and the contributions to the cumulants made by the specific mutation and selection processes occurring in a finite population. In particular, as N is reduced, selection causes a loss in fidelity of the higher order cumulants due to stochastic factors which build up over time.

Negative selection for small rates of site creation

We have been assuming $\frac{x}{\overline{x}} \approx 1$. This assumption is invalid if a majority of organisms, N_0, have no unsatisfied sites and the remaining organisms have just one site. This is the region of selective maintenance of a lack of mutational opportunity sites. In this case a similar analysis to what was applied for negative selection in a sexual population applies,

$$\overline{x} = -\frac{\Gamma_n \log(1+s)}{s} \tag{25}$$

This equation is expected to apply when $\mu_{ss} \ll s$ and in the time it takes negative sites to be cleared new negative sites are unlikely to be created along some linage. That is, when

$$\frac{\log \frac{N}{N_0}}{s} \ll \frac{1}{\Gamma_n}$$

This is certainly true when $\Gamma_n \leq s$.

Positive and negative selection of neutral sites

When $N \ll \frac{1}{s}$ random drift dominates over selection, and the analysis of neutral sites for a sexual population applies,

$$\overline{x} = -\frac{\Gamma_0 \log(1+s)}{\mu_{ss}} \tag{26}$$

Discussion

It is important to be aware of what has not been established. For the asexual case, although a relationship between fitness and variance or skewness has been developed, an expected value for skewness or variance has not been derived. Also an asexual model capable of handling a distribution of fitness effects has not been developed.

Experimental attempts to measure the effects of sex on fitness have represented a mixed bag[15, Table 2]. This is unfortunately understandable. According to equations 1 and 3 log fitness is a linear function of the rate of new true advantageous site creation caused by changes in the environment, Γ_p^*. Biologists today appear to have little understanding of the appropriate value for Γ_p^*, nor how to experimentally control it. Further, experimentally large populations, of say $N \geq 100{,}000$, need to be studied for mean skewness and variance not to largely stochastically wander. The number of generations studied also needs to be large so that average effects occur and so that mutations, with a per site probability, μ_{ss}, of say 10^{-9}, occur in the population.

A model of mutation and selection that incorporates beneficial site creation was presented and analyzed. For sexual populations fitness was completely determined for an arbitrary site creation fitness distribution as a function of the model parameters. For asexual populations fitness was determined, for a single selection coefficient, as a function of the model parameters plus a cumulant of the population fitness distribution.

By incorporating beneficial site creation into mutation-selection models it is possible to determine the mutation rate that maximizes fitness. Too low a mutation rate, and adaptive mutational opportunities take a long time to fix. Too high a mutation rate, and negative selection incurs a heavy cost. The optimal mutation rate for sexual populations is determined to be one for which the fitness losses from positive and negative selection are equal. This suggests that the incorporation of the rate of beneficial site creation offers an important improvement to mutation-selection models.

For asexual populations, selection creates skewness, which causes changes in variance. Variance then drives changes in log fitness. Counteracting the force of selection is mutation, leading to a steady state solution in which the rate of mutation, μ_{ss}, appears in the denominator. Since μ_{ss} is typically very small, mean log fitness may be substantial and negative, making its value sans log minuscule. Sex may create an escape from the large cost that can result.

Acknowledgments

I am very grateful for the time Steven Greidinger, Philip Gerrish, and Benjamin Good spent reviewing early versions of this manuscript and providing me with feedback.

Conflict of interest disclosure

The author declares they have no financial conflicts of interest in relation to the content of this manuscript.

References

[1] H Allen Orr. The population genetics of beneficial mutations. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 365(1544):1195–1201, 2010.

[2] Paul D Sniegowski and Philip J Gerrish. Beneficial mutations and the dynamics of adaptation in asexual populations. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 365(1544):1255–1263, 2010.

[3] Matthew Hartfield and Peter D Keightley. Current hypotheses for the evolution of sex and recombination. *Integrative Zoology*, 7(2):192–209, 2012.

[4] John H Gillespie. *Population genetics: a concise guide*. JHU Press, 2004.

[5] Matt Ridley. *The red queen: sex and the evolution of human nature*. Penguin, 1994.

[6] Stephan Schiffels, Gergely J Szöllősi, Ville Mustonen, and Michael Lässig. Emergent neutrality in adaptive asexual evolution. *Genetics*, 189(4):1361–1375, 2011.

[7] Leo W Beukeboom and Nicolas Perrin. *The evolution of sex determination*. Oxford University Press, 2014.

[8] Motoo Kimura. On the probability of fixation of mutant genes in a population. *Genetics*, 47(6):713, 1962.

[9] Michael M Desai and Daniel S Fisher. Beneficial mutation–selection balance and the effect of linkage on positive selection. *Genetics*, 176(3):1759–1798, 2007.

[10] Motoo Kimura. *The neutral theory of molecular evolution*. Cambridge University Press, 1983.

[11] Michael Lynch. The lower bound to the evolution of mutation rates. *Genome Biology and Evolution*, 3:1107–1118, 2011.

[12] Philip J Gerrish and Paul D Sniegowski. Adding dynamical sufficiency to Fisher's fundamental theorem of natural selection. In *AIP Conference Proceedings: Numerical Analysis and Appied Mathematics ICNAAM 2011*, volume 1389, pages 1260–1262. American Institute of Physics, 2011.

[13] Philip J Gerrish and Paul D Sniegowski. Real time forecasting of near-future evolution. *Journal of the Royal Society Interface*, 9(74):2268–2278, 2012.

[14] Benjamin H Good and Michael M Desai. Fluctuations in fitness distributions and the effects of weak linked selection on sequence evolution. *Theoretical Population Biology*, 85:86–102, 2013.

[15] Lutz Becks and Aneil F Agrawal. The effect of sex on the mean and variance of fitness in facultatively sexual rotifers. *Journal of Evolutionary Biology*, 24(3):656–664, 2011.