

# Evolution of mutation rates in a population

*Michael Anselmi*

## 1 Introduction

Population genetics is a field of biology that concerns itself with the genetic basis of evolution, and differs from many other biological fields in that most of its insights are theoretical and neither observational nor experimental. [1] Just as different branches of mathematics have different central objects of study—homeomorphisms in topology; groups, rings and fields in abstract algebra—so too do the various fields of biology. In population genetics, the primary objects of study are the frequencies and fitnesses of genotypes in populations.

As far as population genetics is concerned, evolution is the change in genotype frequencies in a population over time. Unfortunately, it is impossible to observe these changes directly as the time scale of evolution can be on the order of millions of years. Consequently, we may observe the state of a population at any given point in time, but we cannot directly track how a population evolves. Fortunately all hope is not lost. In order to gain insight into how a population evolves, one constructs mathematical models and studies their behavior, checking to see if the model predicts the state or states in which we expect the population to be under certain circumstances.

In the context of population genetics, the question “how does evolution proceed?” would be expressed as “how do allele frequencies change over time?” In this paper, we attempt to model the interaction of deleterious mutations and variable mutation rates in an idealized population. We would like to gain insight into how mutational forces can steer evolution over large time scales. To this end we begin to develop an iterative model that, when supplied with biologically meaningful parameters, will yield data that reveal the distribution of mutation rates throughout any and all of a population’s equilibrium states. Additionally, we would like this data to suggest a closed-form description of those equilibria.

Why is such a pursuit of interest to us? An example: many strains of bacteria are harmless to humans; in fact, the human gut is teeming with bacteria. However, for certain species of bacteria it is believed that strains which accrue mutations more quickly than the wild-type strain are largely responsible for the emergence of antibiotic resistance in the species. [2] Understanding the conditions necessary for the development of mutator strains may augment our understanding of such emergences.



Figure 1: A snippet of DNA with no base-substitution mutations

## 2 Background

Mutation is the ultimate source of genetic variation. Therefore, understanding the mechanisms driving evolution requires understanding the mechanisms driving mutation. We now make precise what we mean by “mutation” in this paper.

Throughout DNA we find the traditional Watson-Crick bases: adenine, thymine, cytosine, and guanine. Normally, as is depicted in Figure 1, adenine is paired with thymine and cytosine is paired with guanine. However, when we encounter what is known as a single base-substitution (henceforth referred to as a mutation), a single base mutates to another base, resulting in a mismatch. Here we assume that any given mutation leads to exactly one of two possible outcomes. Suppose the thymine in Figure 1 mutates to the guanine in Figure 2. This base pair mismatch may alter the individual’s fitness (reproductive viability). The individual’s fitness could increase, decrease, or remain unchanged. However, since most mutations that do alter fitness do so deleteriously, we assume that each mutation to fitness results in a reduction of fitness. We call such a mutation a type 1 mutation.

DNA-based life has a genetic repair system called mismatch repair that corrects errors like those caused by type 1 and type 2 mutations. However, if we develop a mutation at a locus responsible for the proper functioning of mismatch repair, the performance and reliability of mismatch repair is expected to decrease. This does not directly decrease an individual’s fitness, but it leads to an increased rate of accrual of both type 1 and type 2 mutations in the future. A mutation of this kind is a type 2 mutation. We assume that each type 2 mutation increases an individual’s mutation rate.

In short, a type 1 mutation decreases an individual’s fitness but does not alter his mutation rate. A type 2 mutation does not decrease an individual’s fitness but instead increases his mutation rate.

## 3 Model

Before we can begin developing our model, we must explicitly define our population to be modeled. We must make a number of assumptions about our

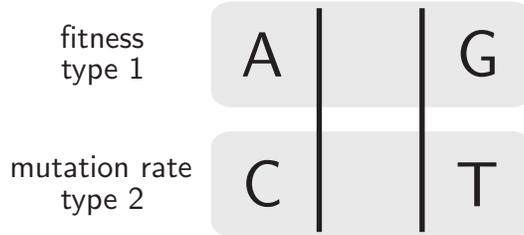


Figure 2: One of each type of mutation has occurred

population in order to develop a reasonably simple model. First, we assume that our population is isolated and is not affected by any forces not accounted for by our model. Second, our population is asexual in the sense that offspring are the product of exactly one parent. Additionally, every individual produces exactly one child and generations do not overlap. Consequently the population size remains fixed. We envision  $N$  individuals reproducing, generating  $N$  offspring, and then dying immediately afterward. Finally we assume that the population size is infinite for reasons that will become apparent later, particularly in section 4.

Every individual in our population is completely specified by two parameters  $x$  and  $y$ , which denote the number of type 1 mutations and the number of type 2 mutations an individual has, respectively. We denote the density of individuals at time  $t$  (generation  $t$ ) with  $x$  number of type 1 mutations and  $y$  number of type 2 mutations by  $D(t)_{x,y}$ . Additionally, we say that an  $(x, y)$  individual has fitness  $w(x)$  and genomic mutation rate  $U(y)$ , where  $w$  is a strictly decreasing function of  $x$  and  $U$  is a strictly increasing function of  $y$ .

An example fitness function  $w$  is  $w(x) = (1 - s)^x$ , where  $w$  maps  $\mathbb{Z}^{\geq}$  (the nonnegative integers) to the half-open interval  $(0, 1]$  and  $s$  is fixed in  $(0, 1)$ . Here, 1 is chosen to be peak fitness, with fitness decreasing toward zero as additional type 1 mutations are accumulated. In population genetics,  $s$  is referred to as the selection coefficient, a value which quantifies the deleterious effect (selective disadvantage) of a single type 1 mutation. This particular fitness function implies multiplicative epistasis of type 1 mutations, by which we mean the following: if two type 1 mutations taken individually each confer a  $1/2$  probability of survival, then taken together they confer a  $(1/2) \cdot (1/2) = 1/4$  probability of survival. That is to say, the effects on the probability of survival of individual type 1 mutations are independent. [1] This function can be modified to be more biologically accurate by mapping to zero after sufficiently many type 1 mutations have been accumulated.

The mutation rate function  $U$  should map  $\mathbb{Z}^{\geq}$  to the interval  $[U_{\min}, \infty)$  or  $[U_{\min}, U_{\max}]$  and should be strictly increasing. We are not yet sure what  $U$  should be, but for the time being we have assumed an exponential growth pattern and chosen  $U$  to be defined by  $U(y) = e^{ky} + c$ , with  $k, c > 0$  fixed. Here we have a minimum mutation rate of  $U_{\min} = 1 + c$ . Biologically speaking, a

positive minimum mutation rate is necessary because no life has or can have a mutation rate of zero. Mathematically speaking, if  $U_{\min}$  were to equal zero, then  $(x, 0)$  would be an absorbing state for all  $x$ .

We wish to calculate  $D(t+1)_{x',y'}$  given that  $D(t)_{x,y}$  is known for all  $x$  and  $y$ .  $D(t+1)_{x',y'}$  is determined in three consecutive steps:

$$D'(t)_{x,y} = D(t)_{x,y} \cdot \frac{w(x)}{\bar{w}}, \quad (1)$$

$$D(t+1)_{x',y'|x,y} = D'(t)_{x,y} \cdot T_{x,y,x',y'}, \quad (2)$$

$$D(t+1)_{x',y'} = \sum_x \sum_y D(t+1)_{x',y'|x,y}. \quad (3)$$

Steps (1) and (2) correspond to natural selection and mutation, respectively. Before a generation reproduces, we must adjust the densities of individuals according to their fitnesses and then normalize by that generation's mean fitness. In the above notation,  $D(t)_{x,y}$  is adjusted to  $D'(t)_{x,y}$  according to fitness. Next, we determine the density of  $(x, y)$  individuals whose offspring are of type  $(x', y')$ . That is, we calculate  $D(t+1)_{x',y'|x,y}$  from  $D'(t)_{x,y}$  and a four-dimensional array  $T$  of transition probabilities, where  $T_{x,y,x',y'}$  denotes the probability that an  $(x, y)$  parent produces an  $(x', y')$  child. Finally, in step (3) we calculate  $D(t+1)_{x',y'}$  by summing over the conditional densities of type  $(x', y')$ .

We describe steps (1) and (2) in detail, beginning with step (1). We start at time 0 and initialize the population by defining various  $D(0)_{x,y}$  as we please. For example, we typically set  $D(0)_{0,0} = 1$  and  $D(0)_{x,y} = 0$  for  $x > 0$  and  $y > 0$ . The term  $w(x)/\bar{w}$  accounts for the effect of type 1 mutations on reproductive output, decreasing  $D(t)_{x,y}$  if  $w(x) < \bar{w}$  and increasing  $D(t)_{x,y}$  if  $w(x) > \bar{w}$ .

Step (2) is easily the most complicated iteration step, with the complexity lying entirely with  $T_{x,y,x',y'}$ . In probability notation, we have that

$$T_{x,y,x',y'} = \Pr\{\text{child } (x', y') \mid \text{parent } (x, y)\}. \quad (4)$$

Assuming an infinite genome size,  $x'$  and  $y'$  become mutually independent, so we rewrite

$$= \Pr\{\text{child } x' \mid \text{parent } (x, y)\} \cdot \Pr\{\text{child } y' \mid \text{parent } (x, y)\}. \quad (5)$$

However, the number of and the rate of arrival of additional type 2 mutations are independent of type 1 mutations, so finally we have

$$T_{x,y,x',y'} = \underbrace{\Pr\{\text{child } x' \mid \text{parent } (x, y)\}}_{P_1} \cdot \underbrace{\Pr\{\text{child } y' \mid \text{parent } y\}}_{P_2}, \quad (6)$$

denoting the first and second terms on the right-hand side of (6) by  $P_1$  and  $P_2$ , respectively. It is here that the assumption of infinite population size is necessary; (2) would not be valid otherwise, as this process would lose its determinicity and become stochastic due to binomial sampling. All that remains

to be done is model  $P_1$  and  $P_2$ . Before continuing, we note that type 1 mutations are assumed not to back-mutate, while type 2 mutations are allowed to back-mutate with a rate independent of the forward mutation rate.

An  $(x', y')$  child of an  $(x, y)$  parent develops exactly  $x' - x$  type 1 mutations not found in the parent, and inherits the remaining  $x$  type 1 mutations from the parent. Let  $k = x' - x$  and let  $W$  be a random variable such that  $\Pr\{W = k\} = \Pr\{\text{child } x' \mid \text{parent } (x, y)\} = P_1$ . Because mutations are rare events and since we are assuming an infinite genome size,  $W$  is well-approximated by a Poisson random variable with parameter  $\lambda$ . Here,  $\lambda = U(y) \cdot f$ , where  $f$  is fixed in  $(0, 1)$  and denotes the proportion of incoming mutations  $U$  that will be of type 1. (Consequently,  $(1 - f)$  denotes the proportion of incoming mutational events of type 2.)  $\lambda$  was so chosen because we require that  $E[W] = U(y) \cdot f$ , and the expected value of any Poisson random variable is its  $\lambda$  parameter. We now have that

$$P_1 = m(k) = \Pr\{W = k\} = \begin{cases} \frac{\lambda^k}{k!} \cdot e^{-\lambda}, & k = 0, 1, 2, \dots \\ 0, & k = -1, -2, -3, \dots \end{cases} \quad (7)$$

Modeling  $P_2$  is a more difficult task. A  $y'$  child of a  $y$  parent accrues  $i$  additional type 2 mutations absent from the parent, and  $j \leq y$  of the parent's type 2 mutations back-mutate in the child such that  $i - j = y' - y$ . Let  $X$  and  $Y$  be random variables denoting the incoming number of forward and backward type 2 mutations, respectively. We have

$$E[X + Y] = U(y) \cdot (1 - f), \quad (8)$$

where  $(1 - f)$  is the proportion of incoming mutations of type 2 (including both forward and back mutation). Thus the random variable  $Z := X - Y$  is such that  $\Pr\{Z = k\} = P_2$ , where  $k = y' - y$ . Unfortunately it is not possible to let  $X$  be a Poisson random variable and  $Y$  be a binomial random variable, for then (8) forces the forward and backward type 2 mutation rates to be mutually dependent. We have not yet identified a model for  $P_2$  suitable for the infinite genome size model. However, should a suitable probability mass function  $n(k)$  of  $Z$  be found, then our model would be complete.

Finally, we halt the iteration of the system once an equilibrium state is encountered, which is when the population's various  $x$  and  $y$  densities no longer change.

## 4 Future work

After choosing a satisfactory model for  $P_2$ , we will then relax the assumption of infinite population size. A finite population size introduces a stochastic element called genetic drift, an unbiased (with respect to selection) dispersive evolutionary force that removes genetic variation from the population; it is mutation's counter-force. [1]

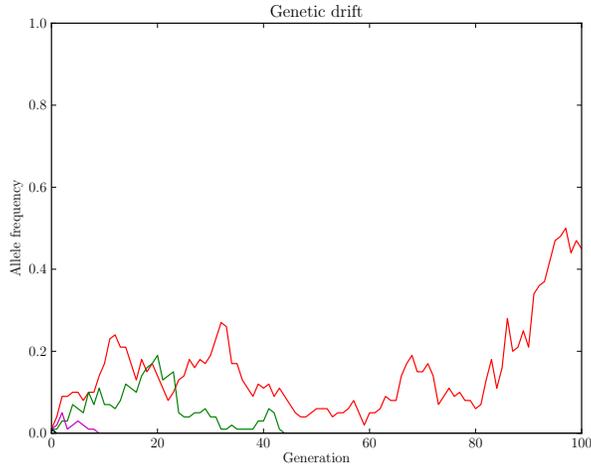


Figure 3: Allele frequency fluctuations due to genetic drift

To conceptualize how genetic drift acts on a population, consider a randomly mating diploid population of  $N$  individuals. For the purposes of genetic drift, it is equivalent to consider this population as a population of  $2N$  alleles of a certain gene. Assuming no evolutionary forces other than genetic drift, the reproductive process is the following:

1. Randomly select one of the  $2N$  alleles from the parent generation.
2. Duplicate the selected allele.
3. Place the duplicate in the new generation. [1]

Suppose our gene under consideration has two mutually exclusive alleles:  $A$  and  $B$ . Further suppose that our population consists of 100 alleles and that the initial frequencies of  $A$  and  $B$  are  $99/100$  and  $1/100$ , respectively. Then there will be a  $(1 - 1/100)^{100} \approx 37\%$  chance that the  $B$  allele will vanish from the population in only one generation. However it is also possible that genetic drift can elevate a new allele to fixation. In Figure 3 we see the effects of drift on the  $B$  allele frequency in several populations, each of which consists of 100 alleles and has  $B$ 's initial frequency set to  $1/100$ .

## 5 Acknowledgments

I thank my advisor, Professor Michael Lynch, for his advice and guidance this summer, along with attempting to teach me a semester's worth of population genetics in only a few weeks' time. I also thank Matthew Ackerman for similar

reasons. Finally, many thanks to Kevin Pilgrim and the National Science Foundation for making this REU possible, and to Amanda McCarty for the many delicious snacks provided throughout the summer.

## References

- [1] John H. Gillespie. *Population Genetics: A Concise Guide*. The Johns Hopkins University Press, 2nd edition, 2004.
- [2] Emmanuel Tannenbaum, Eric Deeds, and Eugene I. Shakhnovich. Equilibrium distribution of mutators in the single fitness peak model. *Physical Review Letters*, 91:138105, 2003.