# A Stochastic Simulation for Evolution in a Changing Environment: Understanding Pyrimethamine Resistance in *Plasmodium falciparum.*

**Ryan Bronson**

Division of Applied Mathematics
Brown University

Spring 2019

Thesis Advisor: Professor Daniel Weinreich
Second Reader: Professor Charles Lawrence

Honors Thesis Submission for Applied Mathematics-Biology Concentration

**Abstract**

As a malaria patient takes and metabolizes weekly doses of an anti-malarial pyrimethamine, there is a periodic change in drug concentration. The growth rate, or "fitness," of a genotype changes with the drug concentration. Previous drug resistance research has considered the time-averaged fitness a suitable statistic, but defining the fitness of a genotype as a dynamic variable can change the outcome of evolution. Generally, there is a fitness trade-off for drug resistance: increased drug resistance comes with a drop in growth rate in the drugless environment. The drug resistant genotype will "win" while drug concentration is high, but if the drug concentration drops, that genotype will "lose" to other, less drug resistant genotypes. This is why changing drug concentration is meaningful. To explore this system, we simulate the evolution of drug resistance of a 16-genotype malaria population within a single human host as the drug concentration changes periodically. The simulation attempts to capture the randomness that exists *in vivo*: it is stochastic in both reproduction and mutation. We also compare several single- and multi-locus positive controls to classical analytic predictions. The simulation is generalized for an array of evolution questions. We expand on classical evolution theory and use the simulation to evaluate our mathematical results for multi-locus, haploid predictions for probability of fixation and heterozygosity. Our investigation of pyrimethamine resistance in malaria suggests that the changing drug concentration in the host does not affect the outcome of evolution because there is no fitness trade-off over clinically relevant, *in vivo* drug concentrations.

# Contents

# 1 Background

## 1.1 Evolution and Population Genetics

Charles Darwin theorized that in a population with diverse heritable traits and finite resources, certain traits that allowed organisms to reproduce more successfully would become more common in the population over time. This natural selection of traits that increase growth rate drives the evolution of species [Darwin, 1859].

### Deterministic Population Genetics

To build a mathematical framework for this phenomenon, we classically define and quantify the growth rate of an individual. In this research, we consider the Malthusian growth rate, or the instantaneous per capita growth rate:

$$r = b - m \ ,$$

where $b$ is the birth rate and $m$ is the mortality rate. We can use this growth rate to model the population over time. The population of $x$ individuals with growth rate $r$ has the following instantaneous rate of change.

$$\frac{dx(t)}{x(t)} = rdt$$

where $x(t) = x_0 \cdot e^{rt}$ models the exponential growth of the population.

If population growth is limited by an environmental carrying capacity $K$, the population experiences logistic growth rather than exponential growth. In this model of population growth,

$$x(t) = \frac{Kx_0 \cdot e^{rt}}{K + x_0 \cdot e^{rt-1}}.$$

The two population models above describe a uniform population, but we are interested in evolutionary population models, which take multiple growth rates into consideration. Because evolution is driven by the competition for scarce resources among individuals with different traits, we also consider the the selection coefficient for genotypes $A$ and $a$:

$$s = r_A - r_a \ .$$

This value determines the success of a trait in a population of alleles $A$ and $a$ because it represents the comparative fitness advantage of allele $A$ over allele $a$. The selection coefficient for two genotypes might also be written as $s_{A,a}$ in systems with more than two alleles or loci under consideration.

**Locus:** a fixed genetic position, a gene or a part of a gene (plural: loci)
**Allele:** a certain variant form at a locus, e.g. a mutation

When a population size is held constant at $N$, we can better examine the competition between two alleles. Such a model is called soft selection: it only considers the replacement of type $a$ individuals with type $A$ individuals. When type $A$ has replaced all of the type $a$ individuals, it has made a selective sweep through the population to achieve fixation.

### Stochastic Population Genetics

In a stochastic model, a beneficial mutation is not guaranteed to fix. It might be lost due to random genetic drift. Genetic drift describes the fluctuation of genotype frequencies in a population due to the random nature of reproduction and mortality. Any one mutant individual has a relative probability of reproducing before death. When the mutation first arises, the single individual that carries it has a nonzero probability of dying before reproducing, even when the mutation is beneficial. If the one mutant fails to reproduce, the mutation is lost. Therefore, while a mutation is at low frequency in the population, its trajectory more susceptible to loss by randomness than the more populous wild type. When the mutant population grows, there is more cushion, so the mutation is less likely to be lost due to the failure of one individual to reproduce. The stochastic treatment of this model considers the probability of fixation of a genotype in a population. The probability of a beneficial mutation reaching fixation in a population of constant size $N$ is

$$P_{fix} = \frac{1 - e^{-2s}}{1 - e^{-2Ns}}$$

where $s$ is the selection coefficient [Kimura, 1964]. When $Ns$ is large, this probability approaches $1 - e^{-2s}$, and when $0 < s \ll 1$, it is approximately $2s$.

Population geneticists have described the "sticky barrier" [Gillespie, 1994] as the minimum mutant frequency beyond which the population is no longer susceptible to loss by genetic drift. Beyong the sticky barrier, the mutation frequency grows logistically according to the deterministic expectation described above. Developing stochastic models is necessary for this population genetics research because of the challenge posed by the sticky barrier to the success of a beneficial mutation.

In the case of a neutral mutation, $s = 0$, but there is still a nonzero probability of fixation in the population due to genetic drift. The probability fixation of a neutral mutation is $\frac{1}{N}$.

### Epistasis in Evolution

Thus far we have only considered single-locus models. When mutations at multiple loci affect the growth rates of individuals in the population, we must examine new levels of complexity. When we examine a population with mutations at multiple loci, there can be more than two genotypes and more than one selection coefficient. In fact, for a system with $L$ loci, there can be $2^L - 1$ distinct selection coefficients to consider. Therefore, our expectations for speed and outcome of evolution are complicated by the introduction of new loci in the presence of epistasis.

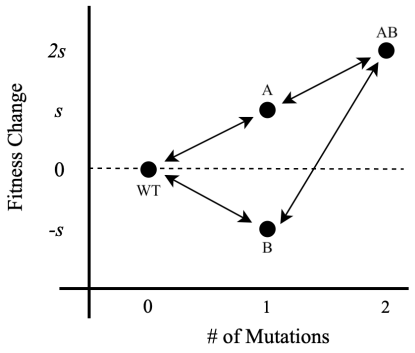


Figure 1: **Epistasis, Two Mutations**

Epistasis is the interaction among genes such that a mutation's contribution to the genotype growth rate is dependent on the genetic background. Within an organism, every mutation can have an unexpected interaction with another mutation or subset of mutations such that the growth rate of a certain genotype might be unpredictable based on the growth rates of its constituent mutations in isolation [Weinreich et al., 2013]. For example, a mutation $B$ that is beneficial in the presence of mutation $A$ could be deleterious without other mutations. The effect of the mutations in combination is not equal to the linear combination of their respective fitness effects alone. In such cases as this, the probability of fixation for mutation $B$ depends on the presence or absence of the other mutations. We describe epistatic such interactions using a "fitness landscape," which represents a combinatorially complete set of growth rates for all possible genotypes.

## 1.2 Drug Resistance

Antimicrobial resistance (AMR) is a growing global health crisis borne out of globalization and the overuse of antimicrobial drugs in human healthcare and animal agriculture. The World Health Organization reported over 23,000 annual deaths due to AMR in the United States, over 25,000 in the European Union, and over 38,000 in Thailand. The WHO predicts that by 2050, antimicrobial resistance will lead to ten million deaths every year [WHO, 2014]. Understanding the evolution of drug resistance, and how certain dosage regimens might inhibit it, might give physicians and public health officials another tool for developing strategies to control and reduce the global burden of disease.

We consider the evolution of drug resistance in a changing drug environment. As the human body metabolizes each in a series of antimicrobial drug doses, there is a periodic fluctuation in the drug concentration. This fluctuation depends on the size and frequency of the doses. The fluctuation in the chemical environment of the microbes induces dynamic selective pressure on the microbial population. Most drug resistance research quantifies fitness for natural selection using constant values. In some cases, researchers assume constant drug concentration (in the human host, this best approximates an IV-drip treatment). Some have calculated the time-averaged mean fitness across a range of drug concentrations. This metric can be sufficient when the drug concentration changes much more quickly than one genotype can sweep through the population. That said, we are interested in the effect of changing drug concentration over time under a stochastic treatment, which could reveal dynamics obscured by previous work in this field [Ogbunugafor et al., 2016, Pou, 2017].

We can measure and estimate different growth rates experimentally. Researchers

have measured the growth rates of organisms at different drug concentrations in order to understand and model the degree of drug resistance exhibited by a certain genotype [Brown et al., 2010]. They determine the drugless growth rate, $R_{00}$, plus growth rates at an array of drug concentrations. Using these data, researchers can estimate the concentration at which replication is cut in half, or $IC_{50}$, by fitting the data to a dose-response curve. $IC_{50}$ is a commonly-used statistic for drug resistance, but it does not give a full picture of how natural selection acts on drug resistance. Because drug concentration in this model is not constant—and because natural selection acts on the current growth rate, not $IC_{50}$—we use the fitted dose-response curve to estimate the growth rate of a genotype at any drug concentration [Ogbunugafor et al., 2016]. It is this fitness value that is actively under selective pressure in the evolution of drug resistance.

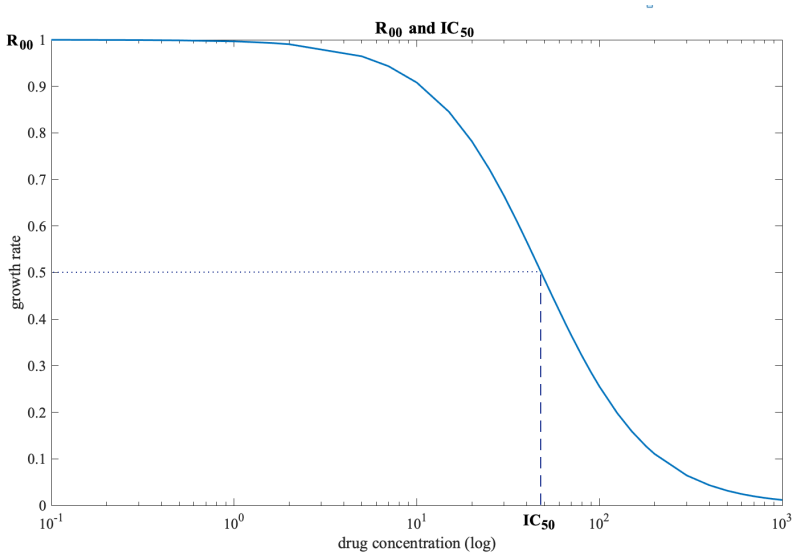


Figure 2: **Normalized Dose-Response Curve:** Growth rate relative to $R_{00}$ by drug concentration. $R_{00} = 1$ when the drug concentration is zero. $IC_{50}$ is the drug concentration at which the growth rate is $\frac{R_{00}}{2}$.

The dynamic drug environment can produce a different outcome of drug resistance evolution than in a constant drug environment. This is possible because there is usually a fitness tradeoff to drug resistance. An organism with high $IC_{50}$ has a new molecular function that allows for drug resistance, likely at a metabolic cost to the organism. Therefore, genotypes with high $IC_{50}$ may have low $R_{00}$, and genotypes with high $R_{00}$ may have low $IC_{50}$. If the population experiences a range of drug concentrations, then the mutations that are favored by natural selection will change periodically. Depending on the dosage size and frequency, the change in drug concentration will affect the "winner" of evolution, or prevent one genotype from "winning" at all. Simulating these dynamics might shed light on dosage strategies that might minimize drug resistance.

## 1.3 Malaria

In 2017, the parasite *Plasmodium falciparum* caused 219 million cases of malaria. While the disease is present over 90 countries, the disease has been endemic in Africa and Southeast Asia for millenia. In fact, 92% of malaria cases occur in Africa [WHO, 2019]. After decades of treating malaria with pyrimethamine, resistance was widespread. Global health officials shifted to artemisinin and artemisinin combination treatment (ACT) in part because resistance to pyrimethamine became unavoidable, but also because artemisinin is inexpensive and highly effective. Artemisinin was even used instead of pyrimethamine in Southeast Asia before pyrimethamine resistance was a major concern. Unfortunately, resistance to artemisinin is now a public health crisis in Southeast Asia, where there were 1.4 billion people at risk of malaria infection in 2015 [WHO, 2016].

Due to the discovery of ACT, pyrimethamine resistance is not a major or growing global health threat. However, there exists an extensive body of research into the molecular biology and epidemiology of this system of resistance. Therefore, our investigation into pyrimethamine resistance is motivated by a rich foundation of previous research and experimental data.

Antifolates like pyrimethamine target dihydrofolate reductase (DHFR), an enzyme involved in DNA synthesis. Lozovsky et al. identified four point mutations (Table 1) in the *dhfr* gene that are associated with resistance to pyrimethamine. The *P. falciparum* pyrimethamine resistance mutational landscape is thus comprised of these four biallelic loci. We refer to the wild-type genotype as 0000. Each 0 represents the lack of mutation at each of the four mutational sites, where a 1 would denote a point mutation. The digits are listed in left $\rightarrow$ right order in the *dhfr* gene (as in Table 1). The triple-mutant 1110 represents the global peak in the drugless fitness landscape. It is also a global peak at all clinically relevant drug concentrations. The quadruple mutant 1111 has a higher $IC_{50}$, but such a drug concentration is never realized in the clinical treatment of malaria.

| Mutation | **N51I** | **C59R** | **S108N** | **I164L** |
|---|---|---|---|---|
| Position | 51 | 59 | 108 | 164 |
| Wild-type residue | asparagine (N) | cysteine | serine | isoleucine |
| Mutant residue | isoleucine (I) | arginine | asparagine | leucine |

Table 1: The leftmost letter is the wildtype amino acid and the rightmost letter is the new amino acid that a mutation at the indicated left $\rightarrow$ right numerical position produces For example, N51I means that the asparagine residue at position 51 has been mutated to isoleucine.

Ogbunugafor et al. investigated the use of $IC_{50}$ as a lone measure of drug resistance in malarial drug resistance research. They found that it failed to predict the most drug resistant genotype at most drug concentrations because the genotype with the highest $IC_{50}$, 1111, is only the most fit at very high drug concentrations. Additionally, using $IC_{50}$ alone did not reveal all plausible mutational trajectories, because it does not reflect the fitness of genotypes at low drug concentrations. These conclusions were developed *in silico*, but align with results from *in vivo* investigation. Ogbunugafor *et al.* argued that if researchers were measuring growth

rates at an array of drug concentrations in order to estimate $IC_{50}$ with a dose-response curve, they could use the dose-response curve to estimate the growth rates at clinically relevant drug concentrations, and that these growth rates could be more informative than $IC_{50}$. This work motivated our interest in a simulation of changing drug concentration over time. Because $IC_{50}$ is one parameter of the dose-response function for growth rates, we use $IC_{50}$ values in our investigation.

In his senior thesis, Sovijja Pou, Applied Math-Biology Class of 2017, investigated evolution of pyrimethamine resistance in a pharmacokinetically realistic environment. He developed a mathematical model using pharmacokinetics for pyrimethamine from Weidekamm et al., fitness data from Brown et al., a growth rates function from Ogbunugafor et al., and a probabilistic method from Weinreich et al. Pou considered the fitness value ranks for the same 16 genotypes of the *dhfr* gene. The rank order of the genotypes was based on their $IC_{50}$ values or their relative growth rates at a certain drug concentration. He investigated how the rank order of the genotypes changed as the drug environment in the model changed. If the rank order changes when drug concentration changes, it suggests that natural selection would favor genotypes in different order of preference depending on the drug concentration.

While Pou's model showed some rank switching after applying a dynamic drug concentration model, it did not affect the plausible pathways to maximal drug resistance previously determined in a constant-environment treatment. This means the evolutionary trajectory did not change, so any rank switching was not consequential for the evolution of drug resistance in this system. Our work expands Pou's concept beyond a mathematical model by developing a stochastic simulation, which gives insight into the effect of the sticky barrier on the evolution of antifolate resistance.

# 2 Methods

## 2.1 Simulation

This stochastic model seeks to simulate evolution in a changing environment. In every iteration, the whole population starts with no mutations (the initial population structure can be changed for different research questions). Each time step is a generation, so for each new generation, $t$, we have an entirely new population made up of the offspring from the previous generation. We assume constant population size, which means there are just as many parents in generation $t$ as there are offspring in generation $t + 1$. First, we determine the environment configuration at the current time step. This requires a time-dependent environment function. Such models can be for drug or substrate concentration, weather, sun exposure, etc. Calculating the environment configuration at a time point $t$ is a deterministic step.

Next, we use another deterministic model which references empirical data for growth rates at different environment configurations to calculate an estimated growth rate for each genotype for the current environment.

We use these growth rates in conjunction with the genotype frequencies in the

previous generation to stochastically generate the pre-mutation population configuration for this generation. Once we have this population array, we determine which mutations occurred, which requires three steps. First, we stochastically determine the number of mutations occurred in the population this generation. The number of mutations is Poisson-distributed with mean equal to the genome-wide, per-individual, per-generation mutation rate. Next, we stochastically determine the source genotype in which each mutation by sampling a multinomial distribution. The probability of a mutation occurring in a source genotype is dependent only on the frequency of that genotype in the population because the genome-wide, per-individual, per-generation mutation rate is the same across all genotypes.

Once we determine the source of each mutation, we distribute the mutations by sampling another multinomial distribution. We stochastically determine to which genotype each source genotype mutated based on the probability of mutation from the source genotype to every other genotype. These probabilities can be calculated by considering any mutational biases in the species. For example, *P. falciparum* has a high AT-content in its genome, so mutations are more likely to result in a substitution to adenine or thymine.

This concludes the generation's operations. We store the population state after mutation and begin the next generation using this population configuration at the next environmental state.

> For a more detailed explanation of the specifics of the functionality and probability of the model, see Appendix A.

## 2.2 Application to Malaria

To apply this simulation to drug resistance evolution in malaria, we need experimentally measured values for the following parameters: the drug concentration function; the growth rates function and fitting constant; $R_{00}$ and $IC_{50}$ for all genotypes; the genome-wide, per-individual, per-generation mutation rate; and a genotype-to-genotype mutation transition matrix.

We use a drug concentration function from Weidekamm et al. that models a weekly dosage regimen for pyrimethamine in humans. They measured the drug concentration in the blood of malaria patients treated with a weekly 25mg dose of pyrimethamine. After taking measurements at several time points in the dosage cycle, they modeled the pharmacokinetics with a 3-step exponential function. As a patient takes compounding weekly doses, the drug is never completely metabolized, so it accumulates in the blood before the system reaches a cyclical equilibrium. This has a small effect on the drug concentration function, but we simplify (as did Pou) in order to minimize computational complexity. Equation 1 is the final drug concentration function. Pyrimethamine concentration spikes in the 2-6 hours following a new dose and then quickly decays as the body metabolizes the drug. This is a 168-hour (7-day) periodic function.

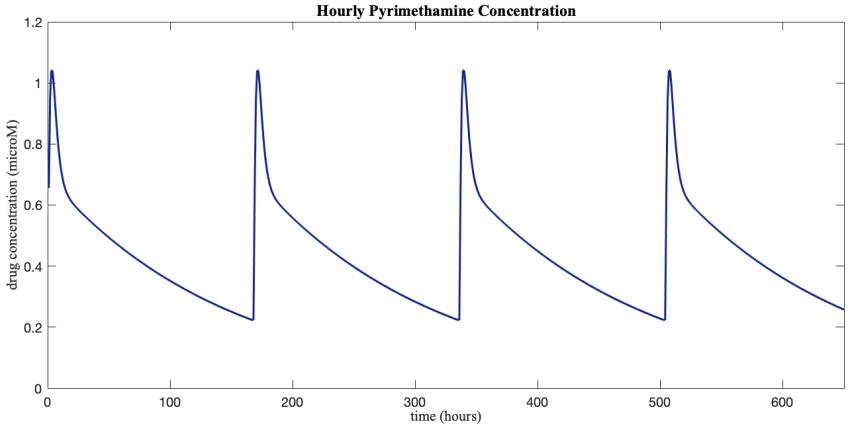


Figure 3: **Hourly Pyrimethamine Concentration:** Drug concentration over time, as modeled by Weidekamm et al. The dosage period is 168 hours, or a week.

$$C(t) = \frac{1}{248710} * (11.193e^{-0.40051(t \bmod 168)} + 0.1723e^{-0.006777(t \bmod 168)} - 11.364e^{-0.4146(t \bmod 168)}) \tag{1}$$

Ogbunugafor et al. used experimental growth rates at a range of drug concentrations to fit a dose-response curve for *P. falciparum* using fitting constant $c = -0.685$.

$$g(x) = \frac{R_{00}}{1 + e^{\frac{IC_{50} - x}{c}}} \tag{2}$$

where $x$ is the current drug concentration and $c$ is an organism-specific fitting constant. Here, current drug concentration and $IC_{50}$ are log-transformed. The drug concentration is one input into the growth rate equation (2), along with $R_{00}$ and $IC_{50}$ values for all 16 genotypes from Lozovsky et al. See Figure 4 for the growth rate of each genotype at a range of drug concentrations.

| **Genotype** | $R_{00}$ | $IC_{50}$ |
|---|---|---|
| 0000 | 1.3978 | 1 |
| 0001 | 1.2749 | 2.9818 |
| 0010 | 1.2275 | 111.5589 |
| 0011 | 0 | $2 * 10^{-5}$ |
| 0100 | 1.3698 | 1.7390 |
| 0101 | 1.3747 | 3.2515 |
| 0110 | 1.3969 | 358.2413 |
| 0111 | 1.2193 | 544.8637 |
| 1000 | 1.1188 | 3.6540 |
| 1001 | 1.1841 | 6.2481 |
| 1010 | 1.3063 | 186.6726 |
| 1011 | 1 | 48.5610 |
| 1100 | 1.2730 | 3.2634 |
| 1101 | 1.2823 | 4.5944 |
| 1110 | 1.4499 | 499.8256 |
| 1111 | 1.2504 | 968.9199 |

Table 2: $R_{00}$ and $IC_{50}$ values for all 16 genotypes for *P. falciparum*. The $R_{00}$ values are relative to slowest-growing viable genotype, 1011, where 0011 is inviable. These values are relative growth rates, so are unitless. The $IC_{50}$ values are normalized such that the wild-type $IC_{50}$ is 1. $IC_{50}$ values are molar concentrations (M). The 0011 $IC_{50}$ was estimated to be 0 but approximated at $2 * 10^{-5}$ so we can log-transform in Equation 2 [Lozovsky et al., 2009]

Paget-McNicol and Saul determined the mutation rate was $1 \times 10^{-9}$ events per *dhfr* gene per generation. We constructed the mutation transition matrix using transition and transversion rates measured by Hamilton et al. It was important to specify such probabilities because *P. falciparum* has a remarkably high AT-content. This means that of all four nucleotides, adenine and thymine comprise almost 80% of bases in coding regions and close to 90% in noncoding regions. Therefore, there is also a mutational bias towards A and T in the *P. falciparum* genome. The transition matrix $\Phi$ gives the probability $\Phi_{i,j}$ of mutation from genotype $i$ to genotype $j$ given that a mutation occurred in genotype $i$. $\Phi$ reflects the mutational bias that has been observed experimentally.

For $R_{00}$ and $IC_{50}$ values, the compounding doses drug concentration function, and the mutation transition matrix, see Appendix A.

## 2.3 Theory Questions

### Probability of Fixation

We build upon the classical model for probability of fixation in a biallelic population to develop a model for the probability of fixation in a three-allele population. In the biallelic model, we have two states: wild-type and mutant. With two states, we have one selection coefficient. In this multi-allelic model there are three states:

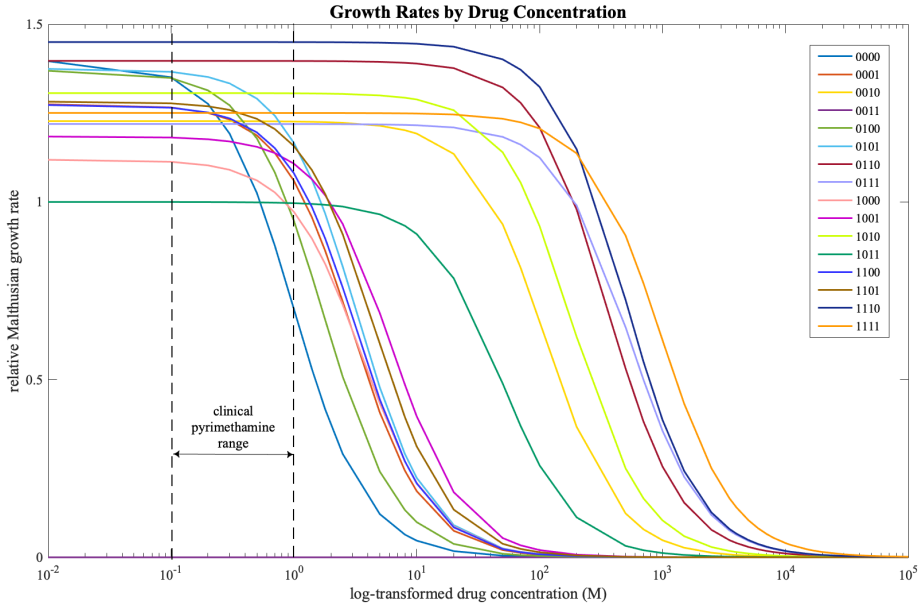


Figure 4: **Growth Rates vs. Drug Concentration:** Normalized dose-response curves for sixteen *P. falciparum* genotypes under a range of drug concentrations, as calculated using Equation 2. Range between dashed lines is the clinically realistic drug concentration range for pyrimethamine in humans.

wild-type and mutants 1 and 2. There are more selection coefficients to consider here. We consider the selection coefficents $s_1$ and $s_2$, which are the fitness advantages of each mutant over the wild-type, respectively. Additionally, $s_2 - s_1$ gives the fitness advantage of mutant 2 over mutant 1. We construct a table of probabilities for different mutation outcomes.

| | mutation lost | mutation fixed |
|---|---|---|
| probability | $1-2s$ | $2s$ |

We extend this table easily for two mutations (three types total). The following table shows the probability of each possible outcome for the population. When both mutations fix, the more beneficial mutation will "win" with probability 1. The more beneficial mutant will always fix in this case because once both mutations have broken through the sticky barrier, they grow deterministically. Therefore, the type with the higher growth rate will sweep through the population quickly. For $s_2 > s_1$, the probability that mutation 1 will "win" is $2s_1 - 4s_1s_2$ and the probability that mutation 2 will "win" is $2s_2$.

| | mutation 1 lost | mutation 1 fixed | row total |
|---|---|---|---|
| mutation 2 lost | $(1-2s_1)(1-2s_2)$ | $(1-2s_2)2s_1$ | $1-2s_2$ |
| mutation 2 fixed | $(1-2s_1)2s_2$ | $4s_1s_2$ | $2s_2$ |
| column total | $1-2s_1$ | $2s_1$ | 1 |

This model assumes small $s$, and can be easily extended for any $s$. We see

that for any $s$, the probability of fixation of the most beneficial mutation in the multi-allelic population is the same as the probability of fixation of a beneficial mutation in the biallelic population.

| | mutation lost | mutation fixed |
|---|---|---|
| probability | $e^{-2s}$ | $1 - e^{-2s}$ |

We extend this table to three types as before. In this case the probability that mutation 1 will "win" is $(1 - e^{-2s_1})(e^{-2s_2})$ and the probability that mutation 2 will "win" is $1 - e^{2s_2}$.

| | mutation 1 lost | mutation 1 fixed | row total |
|---|---|---|---|
| mutation 2 lost | $e^{-2s_1}e^{-2s_2}$ | $e^{-2s_2}(1 - e^{-2s_1})$ | $e^{-2s_2}$ |
| mutation 2 fixed | $e^{-2s_1}(1 - e^{-2s_2})$ | $(1 - e^{-2s_1})(1 - e^{-2s_2})$ | $1 - e^{-2s_2}$ |
| column total | $e^{-2s_1}$ | $1 - e^{-2s_1}$ | 1 |

**Heterozygosity in a Haploid Population**

Heterozygosity is one way to quantify the genetic variation in a population and understand the role of genetic drift on that variation. John H. Gillespie develops a mathematical framework for heterozygosity in a diploid population [Gillespie, 2004]. He assumes any mutation is neutral, where all selection coefficients are equal to zero. Therefore, all change in the population is due to random reproductive and mutation events. We build on Gillespie's framework to derive the following expectation for heterozygosity in a haploid population.

We consider the probability that two randomly selected individuals in the population of size $N$ are different by state, $\mathcal{H}$. We also consider the probability that two randomly selected alleles of different origin are the same state, $\mathcal{G}$, where $\mathcal{H} = 1 - \mathcal{G}$. "Different by origin" alleles are drawn from the population without replacement. We define $\mathcal{H}'$ and $\mathcal{G}'$ as the values of $\mathcal{H}$ and $\mathcal{G}$, respectively, after one round asexual reproduction and random mutation. Gillespie's diploid model instead considers these values after a round of random mating, but the haploid model of genetic drift we consider is dependent only on asexual reproduction and random mutation. Here, the probability of mutation is $\mu$. First, we will derive a haploid model for $\mathcal{G}'$ in Figure 5.

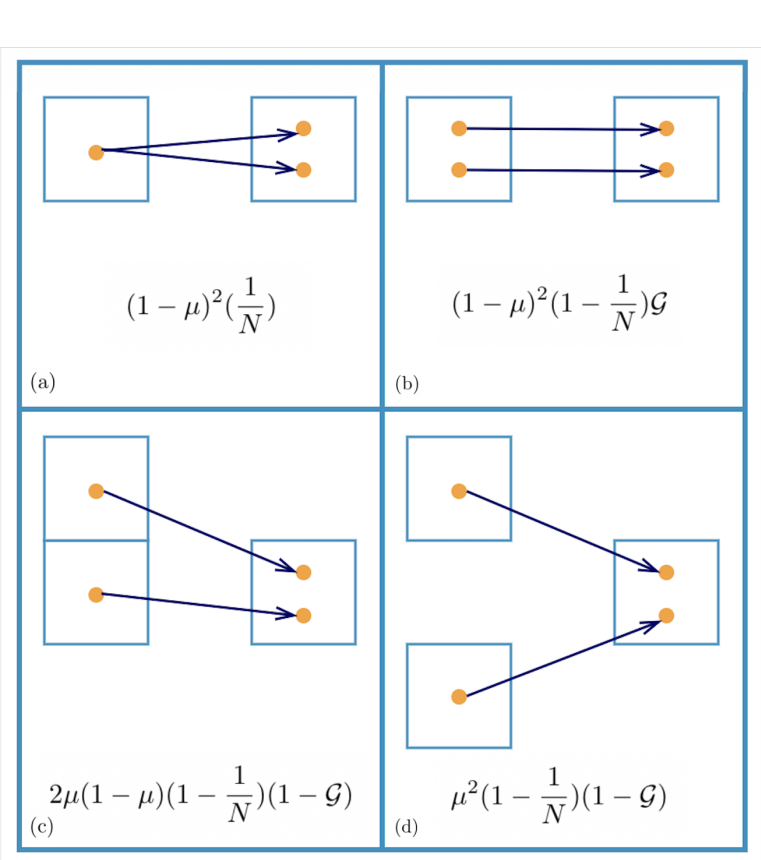


Figure 5: **Derivation of $\mathcal{G}'$:** In each of the four boxes (a-d) above, the leftmost squares are the origin states in the previous generation, $t-1$. The rightmost square is the current state of the randomly selected same-state individuals at time $t$.

There are four cases to consider for the probability that two randomly selected same-state individuals are different by origin following a round of reproduction and mutation. The first (Figure 5a) is the probability that the selected individuals of the same state are the offspring of the same individual in the previous generation, and that no mutation occurred in either reproductive event.

$$P_a = (1-\mu)^2(\frac{1}{N})$$

The second case (Figure 5b) is the probability that the selected individuals of the same state are the offspring of two different, same-state individuals in the previous generation, and that no mutation occured in either reproductive event.

$$P_b = (1-\mu)^2(1-\frac{1}{N})\mathcal{G}$$

The third case (Figure 5c) is the probability that the selected individuals of the same state are the offspring of two different-state individuals in the previous gen-

eration, and that only one mutation occurred in either of the two reproductive events.

$$P_c = 2\mu(1-\mu)(1-\frac{1}{N})(1-\mathcal{G})$$

The final case (Figure 5d) is the probability that the selected individuals of the same state are the offspring of two different-state individuals in the previous generation, and that mutations occurred in both of the two reproductive events.

$$P_d = \mu^2(1-\frac{1}{N})(1-\mathcal{G})$$

$$\mathcal{G}' = P_a + P_b + P_c + P_d$$

We assume very small $\mu$ and very large $N$, so we approximate $(1-\mu)^2$ with $1-2\mu$ and ignore all terms with $\frac{\mu}{N}$, giving

$$\mathcal{G}' = \frac{1}{N}(1-\mathcal{G}) + 2\mu - 4\mu\mathcal{G} + \mathcal{G}$$

We consider $\Delta\mathcal{H}$ the change in $\mathcal{H}$ that occurs in a single generation, where $\Delta\mathcal{H} = \mathcal{H}' - \mathcal{H}$. Again, $\mathcal{H} = 1-\mathcal{G}$ and $\mathcal{H}' = 1-\mathcal{G}'$.

$$\Delta\mathcal{H} = \mathcal{H}(\frac{1}{N} + 4\mu) - 2\mu$$

We are interested in the neutral equilibrium value of heterozygosity in the population, which is where $\Delta\mathcal{H} = 0$.

$$\hat{\mathcal{H}} = \frac{2N\mu}{1+4N\mu}$$

Gillespie similarly derived the diploid model for heterozygosity $\hat{\mathcal{H}}_d$ below. This model will predict greater heterozygosity than our haploid model for all $\mu$ and $N$. This is an intuitive result, as having fewer chromosomes in the popultion results in fewer opportunities for mutation.

$$\hat{\mathcal{H}}_d = \frac{4N\mu}{1+4N\mu} \;>\; \frac{2N\mu}{1+4N\mu}$$

# 3 Results

## 3.1 Simulation results agree with evolution and population genetics theory

### Probability of fixation of a beneficial mutation

We use the classical expectation for probability of fixation in a biallelic population with constant environment as a positive control for the simulation model. In simulations with the parameters in Table 3, we calculated the fixation frequency of the beneficial allele across an array of selection coefficients. Classical modeling finds a probability of fixation of $P_{fix} = 1 - e^{-2s}$. For $s \ll 1$, this is approximately equal to $2s$. Figure 6 shows the mutant population trajectories for 600 iterations

at $s = 0.01$. Trajectories that achieve fixations follow the logistic shape predicted by deterministic and stochastic models. We can visualize the sticky barrier before such trajectories "take off." Below this barrier, many trajectories are lost to or threatened by random drift. These fixation results suggest that this is a high fidelity model with respect to the probability of fixation in a simplified simulation.

| **Fixation Frequency in 1000 Iterations** | | | |
|---|---|---|---|
| | count | count/$(2s * 1000)$ | count/$((1 - e^{-2s}) * 1000)$ |
| $N = 10^5$ | | | |
| $s = 0.005$ | 11 | 1.10 | 1.1055 |
| $s = 0.01$ | 25 | 1.25 | 1.2625 |
| $s = 0.05$ | 79 | 0.79 | 0.8302 |
| $N = 5 * 10^5$ | | | |
| $s = 0.005$ | 12 | 1.20 | 1.2060 |
| $s = 0.01$ | 19 | 0.95 | 0.9595 |
| $s = 0.05$ | 92 | 0.92 | 0.9668 |
| $N = 10^6$ | | | |
| $s = 0.005$ | 10 | 1.00 | 1.0050 |
| $s = 0.01$ | 18 | 0.90 | 0.9090 |
| $s = 0.05$ | 86 | 0.86 | 0.9002 |

Table 3: Fixations of a beneficial mutation starting at one copy in the population. We expect a probability of fixation of $P_{fix} = 1 - e^{-2s}$, which for low $s$, is approximated by $2s$. We expect greater variance for lower $N$.

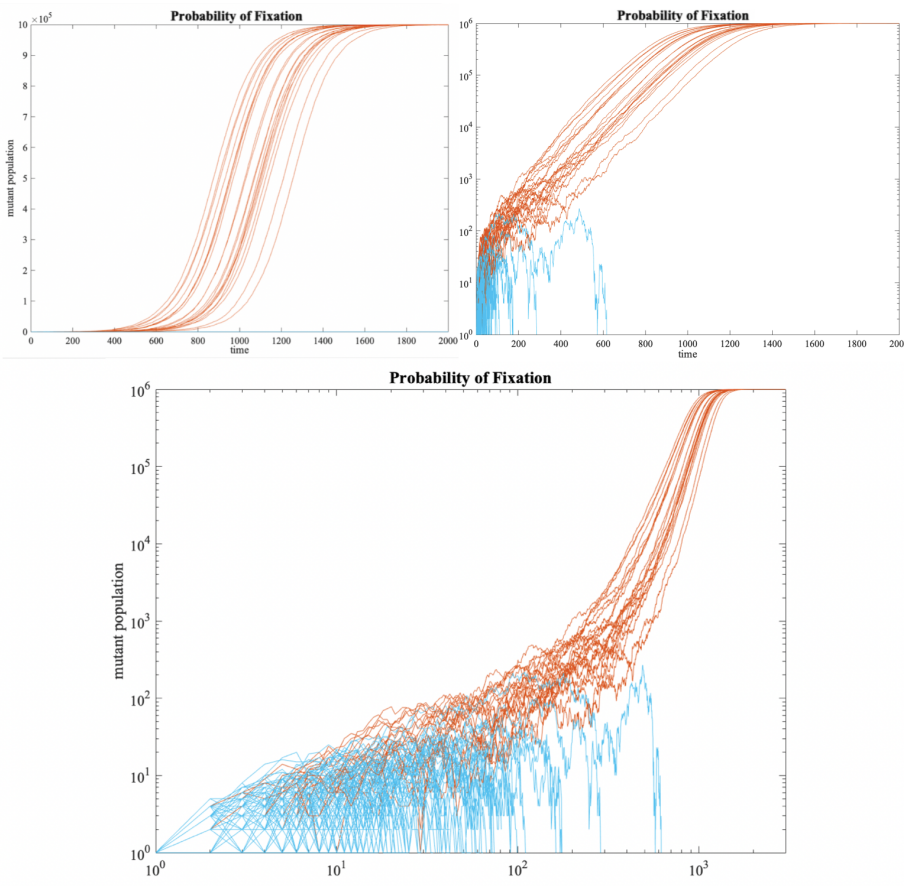


Figure 6: **Probability of Fixation:** Mutant population curves for $s = 0.01$ for 600 iterations with 21 mutation fixations. On the left, linear scales; the right, log population scale and linear time scale; the bottom, log time scales. Successful lineages are in orange, while failed lineages are blue. The left graph shows the logistic curve predicted by the deterministic model for a population sweep. Note how some failed lineages achieve early gains, but are lost to drift below the sticky barrier.

### Probability of fixation of a neutral mutation

We expect a probability of fixation of a single copy of a neutral mutation to be $\frac{1}{N}$. In a population of $N = 10^4$, the average probability of fixation for the neutral mutation was $\frac{0.98}{N}$. As $N$ increases, we should see the probability of fixation approach the classical expectation.

## 3.2 *P. falciparum* resistance to pyrimethamine is not reversible

Pyrimethamine resistance was thought to arise frequently in pathogen populations around the world, but a population-based analysis by Roper et al. revealed that global pyrimethamine resistance can be traced to a single 1110 triple-mutant par-

asite. This genotype then swept through the population and spread around the world. Therefore, our inquiry into the dynamics of the evolution of pyrimethamine resistance is borne out of a rich foundation of previous work, rather than direct clinical applicability. Decades of research produced the fitness and mutation data necessary for our model. The results of this malaria research are not clinically applicable, but the questions drove the structure and functionality of the simulation and provided a positive control for the outcome of simulated evolution.

### Drug resistance landscape

Our investigation did highlight features of the *P. falciparum* pyrimethamine resistance landscape that might inform future drug resistance research. The primary such finding is that this landscape does not feature a consistent fitness trade-off to drug resistance, especially within the clinically realistic pyrimethamine range. This is observable in Figure 4: any genotype with a horizontal dose-response curve (i.e no reduction in growth rate) between the dashed lines has zero fitness trade-off. Rather, selection coefficients acting on these genotypes are constant as the drug concentration changes. Figure 7 shows how growth rates for the wild-type (fitness trade-off) and peak (no trade-off) genotypes change as pyrimethamine concentration changes. We see that not only does the wild-type growth rate drop as drug concentration increases, it also never reaches the peak genotype's constant growth rate.

For comparison, Figure 8 shows how two genotypes with a fitness trade-off might experience rank-switching as drug concentration changes. This figure also shows the time-averaged fitness for each genotype. Note how the average does not reflect the periods of high drug concentration when 1001 grows faster than 1100. Stochastic effects during these periods could result in 1001 "winning" over 1100, a result that a time-averaged fitness model would not predict. While illustrative, this example does not have substantial implications for this study because it does not apply to the genotypes that determine the trajectory to drug resistance. Thus, these fitness trade-offs do not affect the outcome of drug resistance evolution.

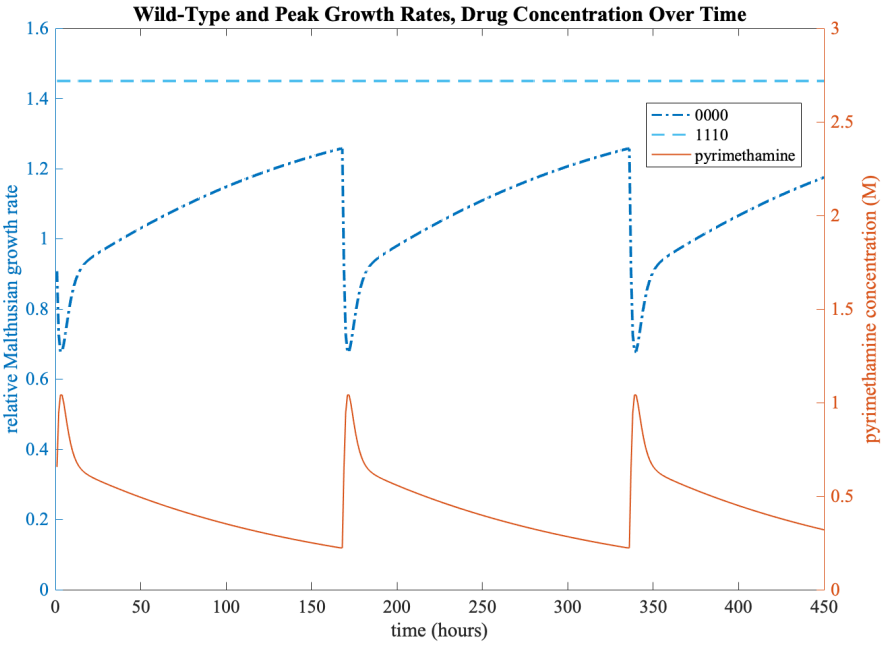


Figure 7: **Growth Rates for Wild-Type and Peak:** Relative growth rates over time for the wild-type and "winner" genotypes as drug concentration changes. The wild-type growth rate is affected by changes in drug concentration, while the peak genotype growth rate remains constant and greater than the wild-type growth rate.

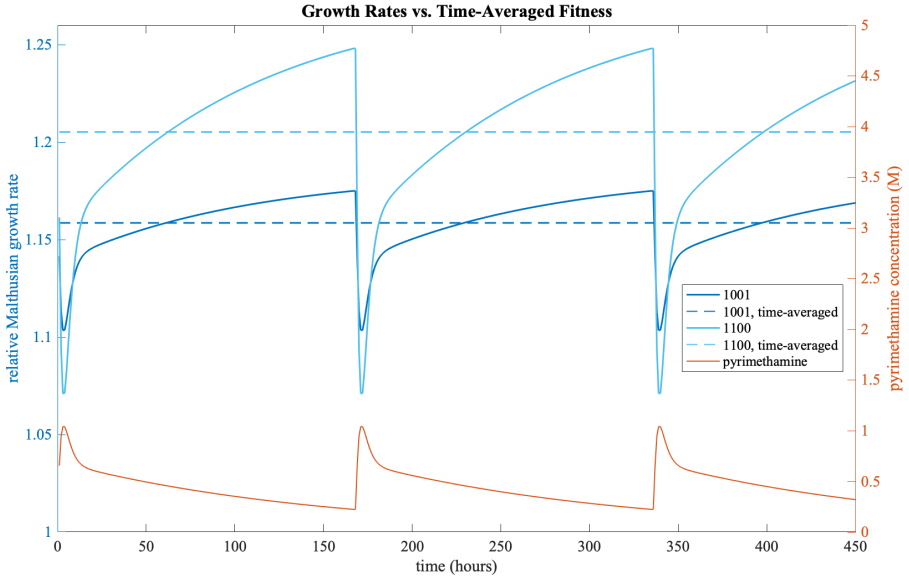


Figure 8: **Growth Rates Over Time vs. Time-Averaged Fitness for 1001 and 1100:** Solid lines: relative growth rates over time for the 1001 and 1100 genotypes as drug concentration changes. Dashed lines: time-averaged fitness for each genotype. The time-averaged fitnesses do not reveal the high-drug concentration periods when 1001 grows faster than 1100.

### Simulations of drug resistance evolution

Simulations of the in-host *P. falciparum* population under pharmacokinetic environmental change shed light on the extent to which changing *in vivo* drug concentration affects the trajectory and outcome of evolution. In two simulations, we examined the population state at each generation. While both simulations had the same expected number of mutations each generation ($N\mu = 0.01$ for both), the simulation with lower $N$ has greater mutation frequency in the population each generation. See Table 4 for the trajectories observed in our simulation and previously documented (Pou 2017).

| **Label** | **Trajectory** |
|---|---|
| 1 | 0000 → 0010 → 0110 → 1110 |
| 2 | 0000 → 0100 → 0110 → 1110 |
| 3 | 0000 → 0010 → 1010 → 1110 |
| 4 | 0000 → 0001 → 0101 |
| 5 | 0000 → 0100 → 0101 |

Table 4: **Evolutionary Trajectories**: Observed trajectories to drug resistance in *P. falciparum*.

The outcome of evolution in this higher-frequency simulation (Figure 10) was almost always the peak genotype 1110. The iterations that resulted in a different

outcome always produced genotype 0101 (Trajectories 4 and 5). This genotype is a local fitness peak, which means that its mutational neighbors (0001, 0100, 0111, and 1101) are less fit at nearly all clinical drug concentrations (Figure 9). Once 0101 sweeps through the population there are no further mutations that increase drug resistance, so the population is "stuck" at a suboptimal genotype.

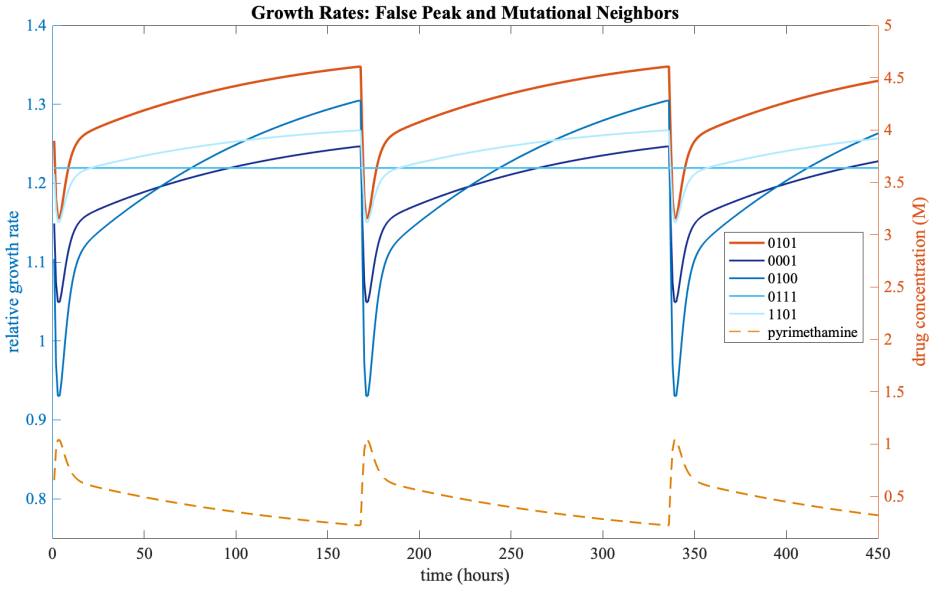


Figure 9: **False Peak and Mutational Neighbors:** Growth rates for the false peak 0101 and its mutational neighbors, 0001, 0100, 0111, and 1101. Note that 0101 has a higher growth rate at nearly all relevant drug concentrations.

The lower-frequency simulation (Figure 11) did not result in the peak genotype as often as the higher-frequency simulation. If the iterations ran long enough, they would eventually reach either the global peak genotype for this system (1110) or the local peak (0101). We can be sure this is the case because as more mutations occur over time, the more-fit genotypes will have repeated opportunities to sweep through the population. The higher-frequency simulation suggested that this is true, and that we would see the same trajectories and outcomes with $N\mu = 0.01$ as the maximum number of generations approaches infinity.

The higher-frequency simulation results more clearly demonstrate the effect of the environment on evolution of drug resistance in this system. We can see periods where a losing genotype begins to regain its frequency in the population (visible environmental change point), but is quickly defeated by the new genotype that was sweeping originally. This occurs often in early drug cycles because the most populous genotypes are those with fitness trade-offs.

### Dosage

All simulations with the standard dosage regimen resulted in a drug resistance peak (either 1010 or 1110). We are interested in a dosage regimen (clinically realistic or

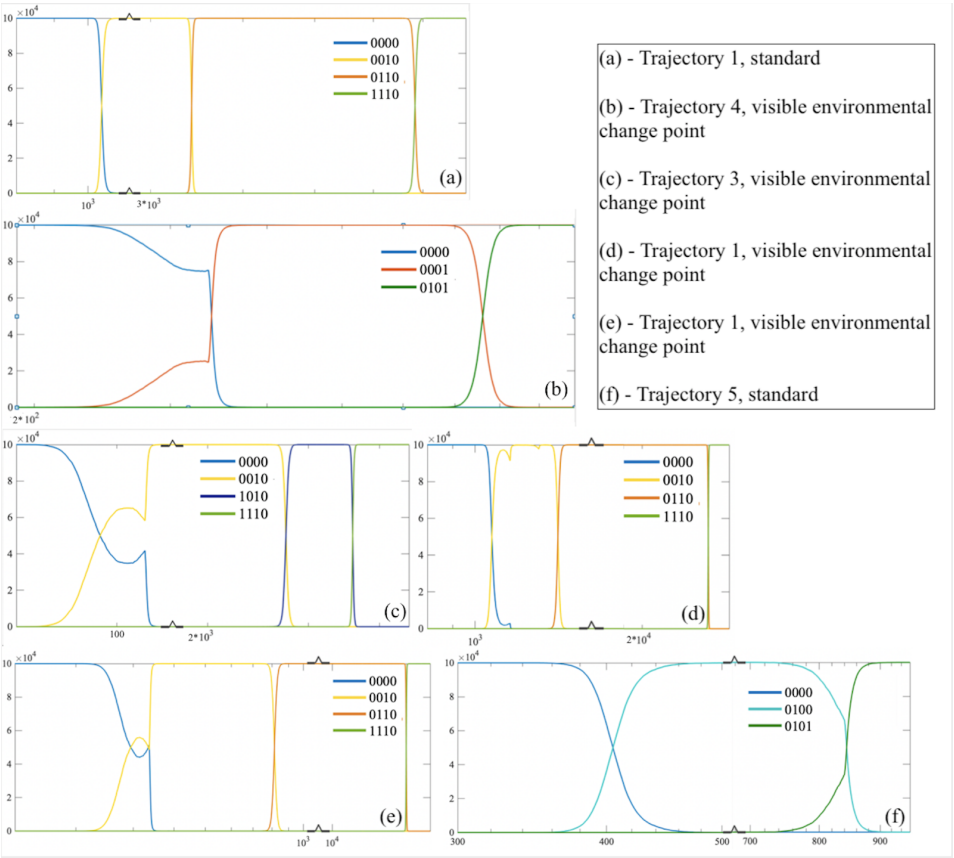


Figure 10: **Simulation of higher mutant frequency:** $N = 10^5$, $\mu = 10^{-7}$. Note that axes have been condensed. Most iterations followed Trajectory 1. Many iterations revealed sharp changes in a genotype's success in the hours immediately following a new dose of pyrimethamine. These iterations are labeled above as having a "visible environmental change point."

not) that will align with the expected time between soft sweeps such that a peak drug-resistant genotype is lost to environmental change. We call such behavior "reversal" of drug resistance. It indicates that altering treatment protocols might prevent the evolution of drug resistance within a patient. reversibility in a fitness landscape also shows how changing the environment function would change the outcome or trajectory of evolution.

In the standard model, we will not see a complete reversal—in which the genotype lost early in the simulation totally regains its fixation in the population—because the dosage regimen is too short. While other dosage regimens might limit the trajectory to peak drug resistance *in silico*, they are likely clinically ineffective or toxic. Additionally, other treatment strategies could be successful *in silico* but unsuccessful *in vivo* because our simulation assumes constant $N$. This means that we could slow the evolution of resistance while allowing the population size to explode within a host, which would likely be associated with dire clinical conse-

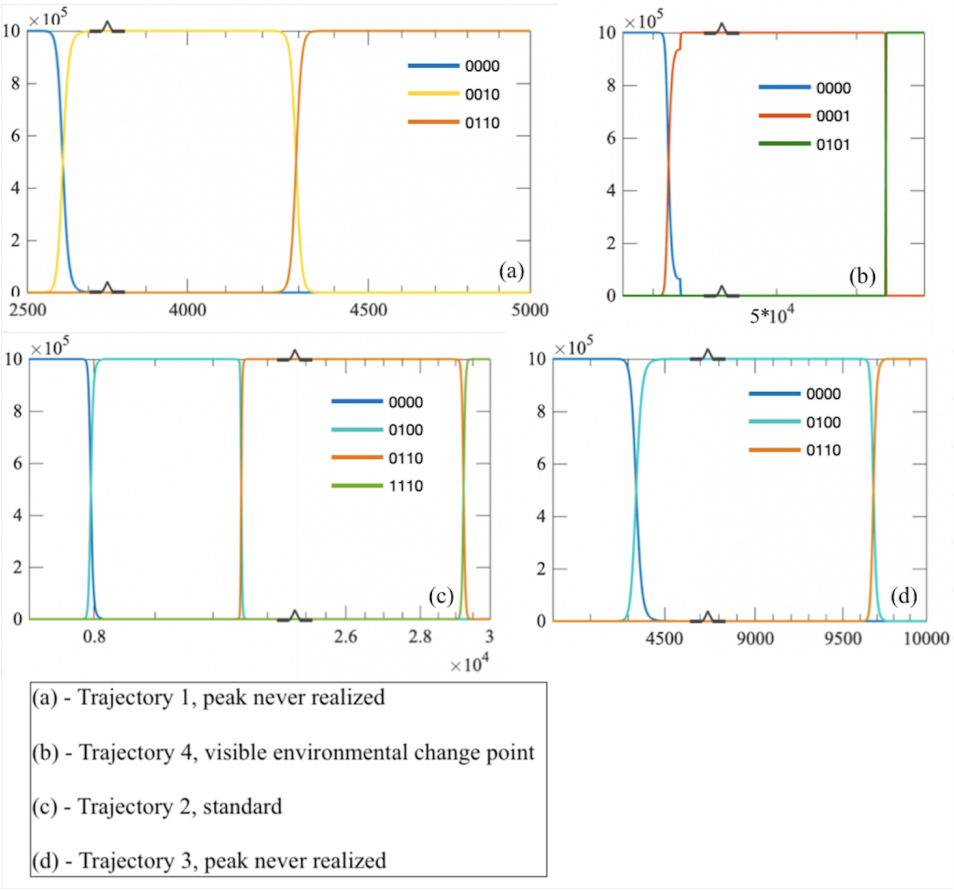


Figure 11: **Simulation of lower mutant frequency:** $N = 10^7$, $\mu = 10^{-9}$. Note that axes have been condensed. Nearly all iterations exhibited standard behavior, and most followed Trajectories 1 and 2.

quences.

Regardless, as a demonstration that the landscape in this case is not reversible, we ran simulations with a different dosage regimens. The regimen in Figure 12 spaces out doses from every 168 hours in the typical treatment to every 400 hours. Under this dosage regimen, the population spends moer time in lower drug concentrations. Notice that even though there is cyclical behavior, this is not the reversal we are interested in because the cycling genotypes are not drug resistance peaks. Additionally, the cycling is always interrupted by more fit genotypes for the environment. Even in iterations where the evolution of drug resistance is slowed, drug resistant mutations will fix in the population according to one of the trajectories in Table 4.

Table 5 gives the number of iterations that resulted in a drug resistance peak for an array of dosage periods. Across dosage periods, nearly all populations arrived at a peak genotype, either the global peak for this system (1110) or the suboptimal

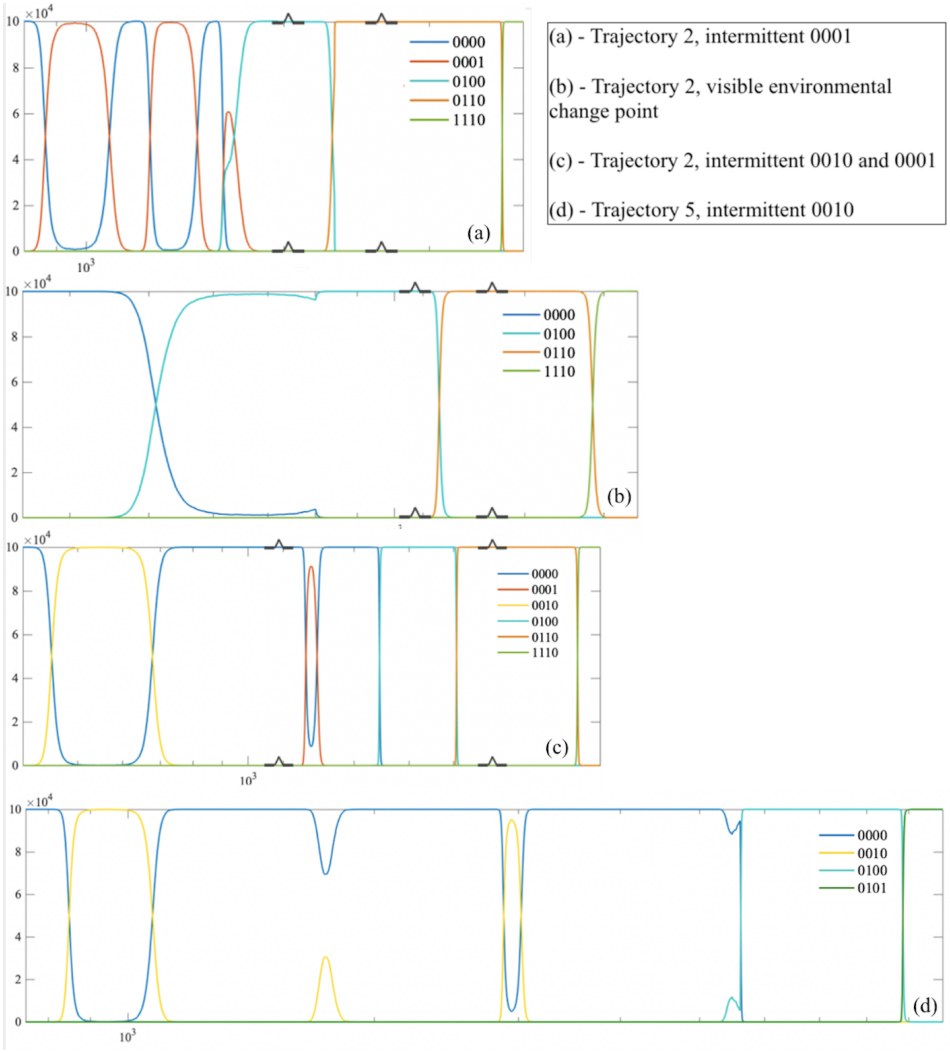


Figure 12: **Simulation of longer dosage periods:** $N = 10^5$, $\mu = 10^{-7}$, dosage period = 400 hours. Note that axes have been condensed. Some iterations followed more standard sweep behavior. Above are the notably different iteration population traces.

local peak (0101), where the population gets stuck. There was no strong trend toward either peak in any simulated dosage period. No dosage regimen that favors genotypes on the mutational trajectory to the drug resistance peak will prevent the evolution of drug resistance. The simulated populations in this work demonstrate that new dosage regimens cannot overcome the nature of the drug resistance landscape for malaria.

| Trajectories by Dosage Period | | | |
|---|---|---|---|
| **Period (hours)** | **Trajectories Observed (out of 100)** | **1110 Fixed** | **0101 Fixed** |
| 100 | [41,21,12,13,5,8] | 77 | 18 |
| 168 | [34,33,5,15,11,2] | 73 | 26 |
| 300 | [10,60,5,7,13,5] | 80 | 20 |
| 400 | [6,56,2,0,14,22] | 75 | 24 |

Table 5: Trajectories observed in 100 iterations each of 4 different dosage periods. Observed trajectory counts listed in the following order: [1, 2, 3, 4, 5, other]. See Table 4 for Trajectories 1-5. Rightmost columns are the count of iterations that ended in each peak. The row sums of the rightmost columns do not always sum to 100 replicates. In some replicates, a peak was not reached in the maximum number of generations ($6 * 10^4$). We expect that all replicates would reach a peak eventually, given enough generations of stochastic mutation and selection.

## 3.3 Probability of fixation for a multi-allele model and heterozygosity for haploid populations

To assess both the fidelity of our simulation and the accuracy of our theory work, we investigated the probability of fixation of three alleles and the heterozygosity of a haploid population in simulation. The results suggest that our math is sound and that the simulation results correctly mimic our model.

### Probability of fixation

We simulated an array of three-allele populations in order to assess the fidelity of our model for the probability of fixation with multiple alleles. In simulations of $10^5$ iterations and $N = 2 * 10^4$, we ran the simulation for four sets of fitness values. Simulation outcomes aligned closely with our the theoretical expectation we derived.

| $(s_1, s_2)$ | **Fixation Freq.** | **Obs/Exp** |
|---|---|---|
| (0.01,0.02) | (0.0190,0.0396) | (1.0000,1.0102) |
| (0.015,0.02) | (0.0283,0.0392) | (0.9965,1.0000) |
| (0.02,0.03) | (0.0372,0.0584) | (1.0081,1.0034) |
| (0.02,0.04) | (0.0362,0.0760) | (1.0000,0.9883) |

Table 6: Fixation frequencies for a population with two mutations. The expectation we use for comparison is that for any $s$. The expected probability of fixation for $s_1$ is $(1-e^{-2s_1})(e^{-2s_2}$. The probability of fixation for $s_2$ is $1 - e^{-2s_2}$.

### Heterozygosity

In a simulation of a population with two neutral genotypes and the parameters below, we calculated experimental heterozygosity at each time step for 20 iterations. We calculate the experimental heterozygosity in the population at each time step

using the following equation.

$$\mathcal{H}_t = 1 - \sum_i p_{i,t}^2$$

where $p_{i,t}$ is the frequency of state $i$ in the population at the current time step.

| $N$ | $t_{max}$ | $\mu$ | initial frequencies |
|---|---|---|---|
| $4 \cdot 10^4$ | $4 \cdot 10^4$ | $4 \cdot 10^{-4}$ | $(\frac{N}{2}, \frac{N}{2})$ |

The average value of experimental heterozygosity across time steps and iterations was $\bar{\mathcal{H}} = 0.4921$. The expectation we developed for this system is $\hat{\mathcal{H}} = 0.4923$.

$$\frac{\bar{\mathcal{H}}}{\hat{\mathcal{H}}} = 0.9996$$

See Figure 13 for experimental heterozygosity plotted on axes with expected heterozygosity.

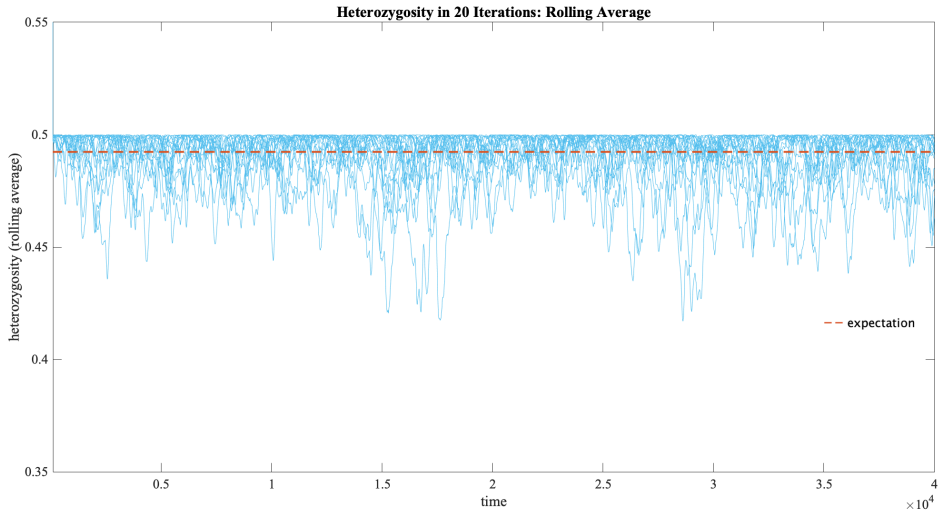


Figure 13: **Heterozygosity in 20 Iterations:** Rolling average of experimental heterozygosity in 20 iterations of $4 \cdot 10^4$ time steps. The sliding window for the rolling average was 50 time steps on either side of the time step in question. Dashed line is expected heterozygosity. Note that the maximum possible value of heterozygosity is 0.5. The average experimental heterozygosity, $\bar{\mathcal{H}}$, was close to our expectation, $\hat{\mathcal{H}}$. $\frac{\bar{\mathcal{H}}}{\hat{\mathcal{H}}} = 0.9996$.

# 4 Discussion

## 4.1 Malarial Drug Resistance

Our simulation only further confirms that pyrimethamine resistance is not reversible, which points to the urgent need for the development of new drugs and

combination treatments as the solution for treating drug-resistant malaria. This research does characterize the features of the fitness landscape that prevents reversibility: no fitness trade-offs across the range of relevant environmental states. Similar drug resistance inquiries might benefit from examining the fitness peaks at different environmental states.

## 4.2 Simulation Applications

This simulation is generalized to be easily applied to any model of genotypic fitness values across a range of environments. Application to drug resistance requires pharmacokinetic models, fitted dose-response curves, mutation rates, and mutation transition probabilities. More generally, this simulation requires growth rate data across an environmental gradient alonside a model for how the environmental gradient changes over time. We asked questions about stochasticity and about reversibility, but the simulation could be modified to ask questions about speed of evolution, evolutionary trajectories, or the effect of different environmental gradient behaviors on the outcome of evolution. See the next section for possible simulation modifications and extensions.

Another interesting application of this simulation is frequency-dependent environmental change. In this case, the environment function, and therefore the growth rates in the population, would be dependent on the frequencies of different genotypes. A biological example of such a system is an organism that consumes substrate from the environment. As it depletes and metabolizes this substrate, it limits the supply of substrate for itself and other organisms. Some genotypes in this species might be better at metabolizing this substrate. Some might grow faster in absence of the substrate. Some might excrete an byproduct substrate after consuming the environment substrate. By modeling these dynamics, one could build a dataset for input into our simulation for evolution in a changing environment.

Finally, this work suggested ways in which a simulation might be helpful in the investigation of evolution and population genetics theory questions. If the simulation's fidelity is confirmed by positive controls from classical evolution and population genetics theory, then extensions to this theory can be examined within the simulation framework. This simulation is especially well-suited for investigations into multi-locus, haploid systems and systems of changing environment and fitness.

## 4.3 Simulation Extensions and Limitations

Many of the limitations of this simulation are due to the population genetics models and assumptions we imposed on the system. Therefore, to overcome these limitations, one need only extend the functionality of the program to account for new population genetics constraints. One such limitation is the type of selection. We use soft selection here because we were interested in selective sweeps and the "winner" of evolution. As such, we sought the size of certain genotypic subpopulations, not the size of the population as a whole. However, in drug resistance research, whether the population will be susceptible to treatment is pertinent. If researchers want to know whether the population is being killed, the simulation can be modified to accommodate a hard selection model with variable total population.

Another limitation of our model is that it only considers discrete generations. This means that generations do not overlap and we only calculate the environmental state once per generation. Thus, the population does not experience constant environmental change within the time-range of a generation. All of the reproduction in a generation occurs under the same selective pressure, which does not perfectly represent evolution under constant environmental change. Another limitation of our generational model is that genotypes with different growth rates have different generation times. While an approximation might be sufficient, a more accurate model would account for the time scale of a generation for each genotype.

This simulation was concerned primarily with the effect of environmental change, so the pharmacokinetic model is deterministic. However, researchers seeking increased stochasticity should sample a normal distribution using the deterministic expectations at each time step. Additionally, in many biological systems, populations are subdivided and experience environmental change spatially. For example, drugs are transported and metabolized at different rates in different parts of the human body. Therefore, organisms in reproducing in different areas of the body will experience different environmental change. A comprehensive extension of this work could investigate the role of subdivided populations in a changing environment on the evolution of drug resistance.

An important simplification in this simulation is the elimination of recombination from the spread of drug resistance. We do not include any recombination functionality, but horizontal recombination is a major factor in the spread of drug resistance, especially in bacterial populations. A recombination extension of this simulation might resemble the mutation functionality, and would likely contribute to computational expense similarly. Another biological simplification made here is the removal of immune system action on the population. If *dhfr* mutations do not affect immune action, then it is likely the immune system would interact with all genotypes similarly. Therefore, our consideration of relative growth rates and constant $N$ would not be affected by the inclusion of the immune system. However, if a biological system under consideration did include differential immune evasion, such modifications could be made in the genotype-specific fitness inputs.

A final limitation is an assumption made in the mutation model. We assume a per generation, per individual, whole-genome mutation rate that is equal across all genetic backgrounds. A final extension to the model could investigate locus-specific mutation rates and incorporate these into the probability of a mutation occuring at each site under consideration.

## 4.4 Statistical Limitations

This work is limited in statistical testing because the stochasticity of this simulation requires running many thousands of iterations in order to observe broader patterns. Therefore, for statistical testing, the sample size is extremely large. Any small deviation in the simulation from the classical expectation would result in extreme $p$-values that would not be highly informative. This limitation is common in modeling research, and is a topic of considerable discussion in this field. Some researchers consider simulation intervals, which reflect the 95% upper and lower limits of the simulation results [Jenness et al., 2016]. Simulation intervals could be

a good approach to statistical testing for these data in the future.

# 5 Acknowledgements

First, I extend my deepest gratitude to my primary advisor, Professor Daniel Weinreich. He has been my mentor and teacher since my first week of college. His courses guided my academic pursuits, but his wisdom and patience as a mentor drove my professional and personal development. Special thanks to Chip Lawrence, my academic advisor, second thesis reader, and professor. I am also grateful to Brandon Ogbunugafor, whose input and support on this project, and as a mentor, were indispensible.

Finally, I would like to thank all of the family and friends who have supported my education and shaped my academic journey.

# Appendices

## Model Specifics

| | |
|---|---|
| $l$ | number of loci |
| $g$ | number of genotypes, $g = 2^l$ |
| $f$ | unit vector representing the probability of an individual of each genotype reproducing in this generation, given the state of the environment, the genotypes' fitness values, and the frequency of the genotype in the population |
| $p$ | vector of $g$ dimensions, each entry of which is a random variable representing the count of that genotype in the population at the current generation |
| $N$ | total population size, constant |
| $\mu$ | per generation, per individual, whole-genome mutation rate |
| $d$ | a unit vector of $g$ dimensions representing the frequency (not count) of each genotype in the current generation: $d = \frac{p_{1:g}}{N}$ |
| $m$ | population-wide count of mutations produced this generation |
| $c$ | vector of $g$ dimensions representing mutation sources |
| $b_i$ | vector of $g$ dimensions representing to which genotype each mutation mutated, for $i = 1, 2, ..., g$ |
| $\Phi$ | $g$-by-$g$ transition matrix of probabilities of mutation from row $i$ to column $j$, rows sum to 1, for $i = j$, $\Phi_{i,j} = 0$ |

### Population State at Each Generation

The first step of each generation is to determine the state of the environment. This is a deterministic step that depends on the environment function under consideration. In this case we use Equation 1. An alternative, more pharmacokinetically accurate model is Equation 3 below.

$$\begin{aligned} C_m(t) = \frac{1}{248710} * (11.193[\frac{1-e^{-0.40051n\tau}}{1-e^{-0.40051\tau}}e^{-0.40051(t \bmod 168)}] \\ +0.1723[\frac{1-e^{-0.006777n\tau}}{1-e^{-0.006777\tau}}e^{-0.006777(t \bmod 168)}] \\ -11.364[\frac{1-e^{-0.4146n\tau}}{1-e^{-0.4146\tau}}e^{-0.4146(t \bmod 168))}] \end{aligned} \tag{3}$$

where $n$ refers to the dose number that is currently being taken. $\tau$ refers to the dosage frequency (168 in this case) while $t$ refers to the total time elapsed (Weidekamm).

Next, we deterministically calculate the growth rate of each genotype. In the malaria application, we use Equation 2. Once we have growth rates, we stochastically determine the pre-mutation population state for this generation. The population state $p$ is a random variable and vector of $g$ dimensions, each entry of which

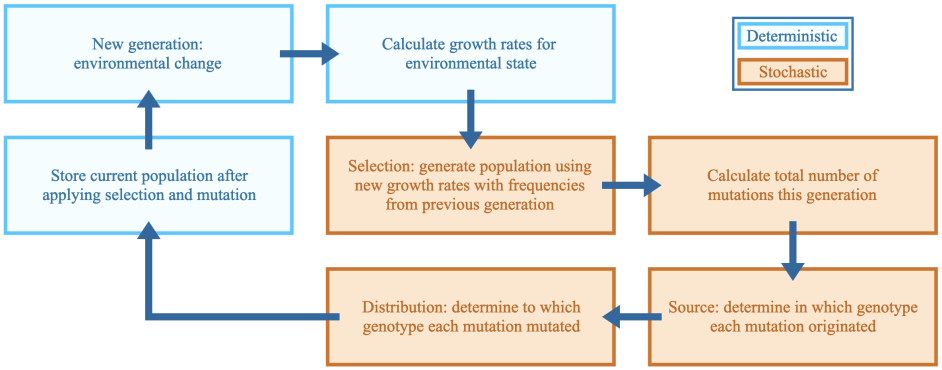


Figure 14: **Simulation Iteration Flowchart:** Each iteration, the simulation calculates the environmental state at the current generation. Growth rates are calculated from the environmental state based on fitness landmark values. There are four successive stochastic steps in the shaded boxes: a) selection is applied using growth rates and frequencies from previous generation, b) the Poisson distribution is sampled for mutation count, c) the source genotype of each mutation is determined using population frequencies, d) the distribution to genotypes of the mutations is determined using the mutation transition matrix. Finally the simulation stores the population trace and repeats.

is a random variable representing the integer count of that genotype in the population in the current generation. We generate $p$ by sampling from the multinomial distribution

$$P(p_{1:g,t} \mid f_{1:g}, N) = Mult(N, f)$$

where $f$ is the unit vector representing the probability of an individual of each genotype reproducing in this generation, given the state of the environment, the genotype fitness values, and the frequency of the genotype in the population. We calculate $f$ by multiplying each entry in $p$ by the corresponding growth rate and normalizing.

$$P(p_{1:g,t} \mid f_{1:g}, N) = N! \prod_{i=1}^{g} \frac{f_i^{p_i}}{p_i!}$$

Because each element in $p$ is a binomially distributed random variable, we have the expected value and variance for each subpopulation $i$

$$E(p_i) = Nf_i, \quad V(p_i) = Nf_i(1 - f_i).$$

## Mutation Count

Next we apply stochastic mutation to the pre-mutation population state for this generation. The mutation count $m$ is a random variable that represents the population-wide integer count of mutations produced in the current generation. We calculate $m$ by sampling from the Poisson distribution with mean $N\mu$

$$P(m \mid N, \mu) = \frac{N\mu^m e^{-N\mu}}{m!}.$$

We can also calculate the expected value and variance of $m$

$$E(m) = N\mu, \quad V(m) = N\mu.$$

## Mutation Sources

We now have the number of mutations in this generation, but we need to determine in which genotypes these mutations originated. These mutation sources are represented by a random variable $c$, a vector of $g$ dimensions. The mutations are distributed based on the genotype frequencies in the population in this generation. We produce $c$ by sampling from the multinomial distribution

$$P(c_{1:g} \mid m, p_{1:g,t}, N) = Mult(m, d)$$

where $d$ is a unit vector of $g$ dimensions representing the frequency (not count) of each genotype in the current generation: $d = \frac{p_{1:g}}{N}$.

$$P(c_{1:g} \mid m, p_{1:g,t}, N) = m! \prod_{i=1}^{g} \frac{(\frac{p_i}{N} c_i)}{c_i!}$$

Because each element in $c$ is a binomially distributed random variable, we have the expected value and variance for each subpopulation $i$

$$E(c_i) = \frac{mp_i}{N}, \quad V(c_i) = \frac{mp_i}{N}(1 - \frac{p_i}{N}).$$

Using the previously stated expected values for $m$ and $p_i$

$$E(c_i) = N\mu f_i, \quad V(c_i) = N\mu * f_i(1 - f_i)(1 - f_i(1 - f_i)).$$

## Mutation Distribution

Using each entry in $c$, we determine to which genotype each mutation mutated. The mutation distribution vector $b_i$ for $i = 1, 2, ..., g$ is a random variable, a vector of neighbors to which each mutant from source genotype $i$ mutated. The sum of the $g$ entries in $b_i$ is equal to $c_i$.The transition matrix, $\Phi$, gives the probability, $\Phi_{i,j}$, of mutation from genotype $i$ to genotype $j$, given that the mutation occurred in genotype $i$. We generate each $b_i$ by sampling fron distinct multinomial distributions

$$P(b_{i,1:g} \mid c_i, \Phi_i) = Mult(c_i, \Phi_i) = c_i! \prod_{j=1}^{g} \frac{\Phi_{i,j}^{b_{i,j}}}{b_{i,j}!}$$

where $\Phi_i$ is the $i$-th row of the mutation transition matrix $\Phi$ below. Because each element in $b_i$ is a binomially distributed random variable, we have the expected value and variance for each subpopulation $j$

$$E(b_{i,j}) = c_i\Phi_{i,j}, \quad V(b_{i,j}) = c_i\Phi_{i,j}(1 - \Phi_{i,j}).$$

Using previously stated expected values for $c_i$ we have an intuitive expected value

$$E(b_{i,j}) = N\mu f_i\Phi_{i,j}.$$

Similarly, but less intuitively,

$$V(b_{i,j}) = N\mu * f_i(1 - f_i)(1 - f_i(1 - f_i))\Phi_{i,j}(1 - \Phi_{i,j}).$$

$$\Phi = \begin{pmatrix}
0 & 0.157 & 0.343 & 0 & 0.343 & 0 & 0 & 0 & 0.157 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0.157 & 0 & 0 & 0.343 & 0 & 0.343 & 0 & 0 & 0 & 0.157 & 0 & 0 & 0 & 0 & 0 & 0 \\
0.343 & 0 & 0 & 0.157 & 0 & 0 & 0.343 & 0 & 0 & 0 & 0.157 & 0 & 0 & 0 & 0 & 0 \\
0 & 0.343 & 0.157 & 0 & 0 & 0 & 0 & 0.343 & 0 & 0 & 0 & 0.157 & 0 & 0 & 0 & 0 \\
0.343 & 0 & 0 & 0 & 0 & 0.157 & 0.343 & 0 & 0 & 0 & 0 & 0 & 0.157 & 0 & 0 & 0 \\
0 & 0.343 & 0 & 0 & 0.157 & 0 & 0 & 0.343 & 0 & 0 & 0 & 0 & 0 & 0.157 & 0 & 0 \\
0 & 0 & 0.343 & 0 & 0.343 & 0 & 0 & 0.157 & 0 & 0 & 0 & 0 & 0 & 0 & 0.157 & 0 \\
0 & 0 & 0 & 0.343 & 0 & 0.343 & 0.157 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0.157 \\
0.157 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0.157 & 0.343 & 0 & 0.343 & 0 & 0 & 0 \\
0 & 0.157 & 0 & 0 & 0 & 0 & 0 & 0 & 0.157 & 0 & 0 & 0.343 & 0 & 0.343 & 0 & 0 \\
0 & 0 & 0.157 & 0 & 0 & 0 & 0 & 0 & 0.343 & 0 & 0 & 0.157 & 0 & 0 & 0.343 & 0 \\
0 & 0 & 0 & 0.157 & 0 & 0 & 0 & 0 & 0 & 0.343 & 0.157 & 0 & 0 & 0 & 0 & 0.343 \\
0 & 0 & 0 & 0 & 0.157 & 0 & 0 & 0 & 0.343 & 0 & 0 & 0 & 0 & 0.157 & 0.343 & 0 \\
0 & 0 & 0 & 0 & 0 & 0.157 & 0 & 0 & 0 & 0.343 & 0 & 0 & 0.157 & 0 & 0 & 0.343 \\
0 & 0 & 0 & 0 & 0 & 0 & 0.157 & 0 & 0 & 0 & 0.343 & 0 & 0.343 & 0 & 0 & 0.157 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0.157 & 0 & 0 & 0 & 0.343 & 0 & 0.343 & 0.157 & 0
\end{pmatrix}$$

# References

[Bauer et al., 2017] Bauer, M., Graf, I. R., Ngampruetikorn, V., Stephens, G. J., and Frey, E. (2017). Exploiting ecology in drug pulse sequences in favour of population reduction. *PLOS Computational Biology*, 13(9):e1005747.

[Brown et al., 2010] Brown, K. M., Costanzo, M. S., Xu, W., Roy, S., Lozovsky, E. R., and Hartl, D. L. (2010). Compensatory mutations restore fitness during the evolution of dihydrofolate reductase. *Molecular Biology and Evolution*, 27(12):2682–2690.

[Costanzo et al., 2011] Costanzo, M. S., Brown, K. M., and Hartl, D. L. (2011). Fitness Trade-Offs in the Evolution of Dihydrofolate Reductase and Drug Resistance in Plasmodium falciparum. *PLOS ONE*, 6(5):e19636.

[Darwin, 1859] Darwin, C. (1859). *On the Origin of Species*. John Murray, London, United Kingdom.

[Gatton et al., 2004] Gatton, M. L., Martin, L. B., and Cheng, Q. (2004). Evolution of Resistance to Sulfadoxine-Pyrimethamine in Plasmodium falciparum. *Antimicrobial Agents and Chemotherapy*, 48(6):2116–2123.

[Gillespie, 1994] Gillespie, J. H. (1994). *The Causes of Molecular Evolution*. Oxford University Press. Google-Books-ID: 257cnXAoREwC.

[Gillespie, 2004] Gillespie, J. H. (2004). *Population Genetics: A Concise Guide*. JHUP, 2nd edition edition.

[Hamilton et al., 2017] Hamilton, W. L., Claessens, A., Otto, T. D., Kekre, M., Fairhurst, R. M., Rayner, J. C., and Kwiatkowski, D. (2017). Extreme mutation bias and high AT content in Plasmodium falciparum. *Nucleic Acids Research*, 45(4):1889–1901.

[Heinberg and Kirkman, 2015] Heinberg, A. and Kirkman, L. (2015). The molecular basis of antifolate resistance in Plasmodium falciparum: looking beyond point mutations. *Annals of the New York Academy of Sciences*, 1342(1):10–18.

[Jenness et al., 2016] Jenness, S. M., Goodreau, S. M., Rosenberg, E., Beylerian, E. N., Hoover, K. W., Smith, D. K., and Sullivan, P. (2016). Impact of the Centers for Disease Control's HIV Preexposure Prophylaxis Guidelines for Men Who Have Sex With Men in the United States. *The Journal of Infectious Diseases*, 214(12):1800–1807.

[Kimura, 1964] Kimura, M. (1964). Diffusion Models in Population Genetics. *Journal of Applied Probability*, 1(2):177–232.

[Lozovsky et al., 2009] Lozovsky, E. R., Chookajorn, T., Brown, K. M., Imwong, M., Shaw, P. J., Kamchonwongpaisan, S., Neafsey, D. E., Weinreich, D. M., and Hartl, D. L. (2009). Stepwise acquisition of pyrimethamine resistance in the malaria parasite. *Proceedings of the National Academy of Sciences*, 106(29):12025–12030.

[Ogbunugafor and Eppstein, 2017] Ogbunugafor, C. B. and Eppstein, M. J. (2017). Competition along trajectories governs adaptation rates towards antimicrobial resistance. *Nature Ecology & Evolution*, 1(1):0007.

[Ogbunugafor et al., 2016] Ogbunugafor, C. B., Wylie, C. S., Diakite, I., Weinreich, D. M., and Hartl, D. L. (2016). Adaptive Landscape by Environment Interactions Dictate Evolutionary Dynamics in Models of Drug Resistance. *PLOS Computational Biology*, 12(1):e1004710.

[Paget-Mcnicol and Saul, 2001] Paget-Mcnicol, S. and Saul, A. (2001). Mutation rates in the dihydrofolate reductase gene of Plasmodium falciparum. *Parasitology*, 122(5):497–505.

[Pou, 2017] Pou, S. (2017). Modeling the Evolution of Antimalarial Drug Resistance for Clinically Realistic Pharmacokinetics. page 28.

[Roper et al., 2003] Roper, C., Pearce, R., Bredenkamp, B., Gumede, J., Drakeley, C., Mosha, F., Chandramohan, D., and Sharp, B. (2003). Antifolate antimalarial resistance in southeast Africa: a population-based analysis. *The Lancet*, 361(9364):1174–1181.

[Sibley et al., 2001] Sibley, C. H., Hyde, J. E., Sims, P. F. G., Plowe, C. V., Kublin, J. G., Mberu, E. K., Cowman, A. F., Winstanley, P. A., Watkins, W. M., and Nzila, A. M. (2001). Pyrimethamine–sulfadoxine resistance in Plasmodium falciparum: what next? *Trends in Parasitology*, 17(12):582–588.

[Weidekamm et al., 1982] Weidekamm, E., Plozza-Nottebrock, H., Forgo, I., and Dubach, U. C. (1982). Plasma concentrations of pyrimethamine and sulfadoxine and evaluation of pharmaco-kinetic data by computerized curve fitting. *Bulletin of the World Health Organization*, 60(1):115–122.

[Weinreich and Chao, 2005] Weinreich, D. M. and Chao, L. (2005). Rapid Evolutionary Escape by Large Populations from Local Fitness Peaks Is Likely in Nature. *Evolution*, 59(6):1175–1182.

[Weinreich et al., 2006] Weinreich, D. M., Delaney, N. F., DePristo, M. A., and Hartl, D. L. (2006). Darwinian Evolution Can Follow Only Very Few Mutational Paths to Fitter Proteins. *Science*, 312(5770):111–114.

[Weinreich et al., 2013] Weinreich, D. M., Lan, Y., Wylie, C. S., and Heckendorn, R. B. (2013). Should evolutionary geneticists worry about higher-order epistasis? *Current Opinion in Genetics & Development*, 23(6):700–707.

[WHO, 2014] WHO (2014). WHO | Antimicrobial resistance: global report on surveillance 2014.

[WHO, 2016] WHO (2016). World Health Organization, Data and statistics.

[WHO, 2019] WHO (2019). Fact sheet about Malaria.