Fitness Estimation in Models of Genetic Evolution of Asexual Populations

Sergey S. Sarkisov II, Ilya Timofeyev, Robert Azencott

November 29, 2018

Abstract

In this paper we develop a systematic methodology for estimating the fitness of genotypes from observational data describing the long-term evolution of bacterial populations. In particular, we develop a numerical approach for estimating genotypes' fitnesses in locked-box models describing bacterial evolution in experimental setups similar to the celebrated Lenski experiment. Our approach is based on a nonlinear least squares optimization procedure which takes into account both the first and the second mutation events. The methodology developed in this paper is confirmed with numerical simulations with realistic parameter values.

AMS Classification: 60J20, 92D15

1 Introduction

Studies of evolutionary dynamics, including bacterial populations, received a considerable amount of attention in recent decades starting with the celebrated Lenski experiment [15, 26]. Since then other groups contributed to experimental studies of Escherichia coli (see e.g. [3, 10, 2, 16, 6, 11, 4]). The experimental setup usually consists of multiple populations processed in parallel. These populations undergo daily growth followed by daily selections of fixed-size samples. Initially, all populations are identical and consist only of the ancestor genotype. Since evolutionary experiments are carried out for a long time (years), different random mutations occur in different populations. Therefore, the genetic composition of different populations can differ drastically over time. The main goal of such experiments is to understand the main features of evolutionary dynamics, including the rate of evolutionary change, likelihood of emergence of a new genotype, fitness landscape, etc.

The Lenski Escherichia coli long-term evolution experiment provides a considerable insight and evidence for the mechanisms of growth and adaptation of asexual organisms (e.g. [15, 26, 8]). Analysis of the experimental data shed some light upon the relative growth rates of stronger mutants and their ancestors. However, the Lenski experiment primarily focuses on the overall population adaptation [1, 27], while our goal is to develop techniques for estimating individual selective advantages for all genotypes.

The adaptive evolution of bacterial populations is driven by the emergence of positive mutations and their spread due to natural selection. Various mathematical models of this evolution process have been developed (see e.g. [12, 25]), but estimation of parameters in these models remains one of the key issues. The evolutionary dynamics of bacterial populations depends on two main parameters - selective advantages and mutation rates [10, 17, 14]. In this work we assume that the mutation rates are known and address estimation of selective advantages.

Several previous works addressed similar questions. For instance, one study involved both experimental and theoretical analyses of evolving viral populations [21, 20]. It has been able to directly
predict adaptation models, including the distribution of beneficial mutation effects. However, this work was limited to the viruses, because of their small genome size, high mutation rate, and large mutation effects. By contrast, in bacterial populations, direct estimates of mutational parameters is much more complex because these parameters are usually estimated indirectly by inference from observable markers. A recent study focused on bacterial populations where a single mutant type emerges and overtakes the population. Mathematical derivations and model simulations were combined to derive accurate estimates for the mutation rates and selective advantages of the main emerging beneficial mutation. In the approach developed only the last and most significant divergences were applied to obtain parameter estimates. In another study on parameter estimation, only the first divergences were used to perform generalized regressions. In contrast with these works, we develop more realistic formulas involving occurrence of multiple mutant types.

A typical approach in E. coli evolutionary experiments is to observe many populations simultaneously evolving in parallel (see e.g. ). Each population consists entirely of the ancestor type cells initially. In order to be able to observe mutation events, half of the cells are colored via a particular marker in each population (e.g. Ara$^+$ marker in E. coli experiments in ), whereas the other half are colored via a different marker (e.g. Ara$^-$). In E. coli evolutionary experiments Ara$^+$ and Ara$^-$ markers correspond to growing white and red cell colonies, respectively. Initially, red and white colonies coexist at the rate 50/50 in all populations. No cell dynamics is altered after applying this coloration. These asexual microorganisms reproduce via binary fission, and the genetic markers are passed down from parents to offspring. Initially these population consist of N cells, with typical values of $N = 50,000,\ldots,100,000$. The populations are allowed to grow for 24 hours, and then a sub-population of approximate size $N$ is extracted by dilution and transferred to a new culture of fresh growth medium. This transfer step is repeated daily for all populations. The white and red cell frequencies are estimated periodically by plating cells on indicator media while visual counting errors are neglected. Therefore, after the initialization step (initialize all populations with the identical ancestor genotype colored 50/50 with the Ara$^+$ and Ara$^-$ markers), the experimental process of growth, mutations, and dilution is repeated for many days.

A mutation occurs if one of the colors significantly exceeds 50% in a particular population. With daily observations, the whole trajectory for one of the colors can be recorded. Since there are multiple populations undergoing evolutionary dynamics, combining all population data results in an ensemble of trajectories. In this paper we develop an optimization procedure which utilizes this ensemble data to estimate the selective advantages of all genotypes in the population. In particular, in section we describe the stochastic Markov chain model which describes the evolutionary dynamics of bacterial populations. In section we derive expressions for the most-likely evolutionary trajectory conditioned on the occurrence of one- and two-mutation events. In section we describe the setup for our numerical experiments to generate synthetic data for an ensemble of evolutionary trajectories. In section we define an optimization procedure for estimating the genotypes’ selective advantages and in section we present numerical results.

2 Stochastic Model

One of the most popular mathematical models describing the biological experiments outlined in the previous paragraph are Markov-chain “locked-box” models. In these models the mutation is separated from growth, and the biological experiment is modeled by the repeating sequence $\text{growth} \Rightarrow \text{mutations} \Rightarrow \text{dilution} \Rightarrow \ldots$ To construct the mathematical model, we introduce the set of genotype frequencies, $\{H_j(t)\}$, where $t \in \{1,2,\ldots,k\}$ is the discrete time measured in days and $j \in \{1,2,3,\ldots,g\}$ is the genotype index. Random variables $H_j(t)$ describe the daily
evolution of the population of \( g \) genotypes (1 is the ancestor genotype and the other \((g - 1)\) are mutants) and, thus, genotypes’ frequencies obey the relationship \( \sum_{j=1}^{g} H_j(t) = 1 \). Next, we describe the three phases in more detail.

2.1 Growth

In the beginning of any day, \( t \), the population consists of a total of \( N \) cells. The population can include various sub-populations of sizes \( N_j(t) = N \times H_j(t) \), so that \( N = \sum_{j=1}^{g} N_j(t) \). The daily growth phase is modeled by a deterministic exponential growth in which cell divisions occur at fixed growth rates \( \rho_j \) per corresponding genotypes \( j \). Hence, after one day, the number of cells \( N_j(t) \) of each sub-population grows by the corresponding growth factor \( F_j = \exp(\rho_j) \), so that the grown subpopulations are of sizes \( S_j(t) = F_j \times N_j(t) \) and the net population at the end of day \( t \) is of the size \( S(t) = \sum_{j=1}^{g} S_j(t) \). Without loss of generality, assume that genotypes are ordered with respect to their fitnesses as \( F_1 < F_2 \cdots < F_{g-1} < F_g \). The selective advantage \( s_j \) for each genotype is defined as \( F_j = F_1^{(1+s_j)} \) and can be easily computed as

\[
    s_j = \frac{\log(F_j)}{\log(F_1)} - 1. \tag{1}
\]

In some models (see e.g. [7, 28]) it is assumed that the amount of nutrients is limited during the day and the population can only grow to a fixed size, \( D \). However, in our model we assume that the population growth is not bounded by any external factors.

2.2 Mutation

After growth, the population is typically several orders of magnitude larger than the ancestor population, and even for small mutation rates of the order of \( 10^{-6} \), a non-zero number of emergent mutants can be expected. Here we consider only advantageous mutations, since deleterious mutations are typically quickly eliminated from the population due to a lower selective advantage. We assume that mutations are independent random events, occurring with equal mutation rates \( \mu \) for \( i = 1 \ldots (g-1) \). Here \( \mu \) represents the probability of mutation; note that it does not specify the genotype after the mutation. The formalism presented here can be easily generalized for different mutation rates \( \mu_i \) which represents the probability of mutation of genotype genotype \( i \). The only restriction for mutation rates is that they are small and are of the same order, i.e. \( \mu_i = O(\mu) \) for \( i = 1 \ldots (g-1) \).

We introduce the transition matrix \( P = \{p_{ij}\} \) describing the probability that mutants from the genotype-\( i \) subpopulation become of genotype \( j \). The mutations are modeled by the Poisson distribution with means

\[
    \lambda_{ij} = \mathbb{E}[\eta_{ij}] = p_{ij} \times \mu \times S_i \quad \text{such that} \quad Pr(\eta_{ij} = k_{ij}) = \frac{\lambda_{ij}^{k_{ij}} e^{-\lambda_{ij}}}{k_{ij}!} \tag{2}
\]

with \( \eta_{ij} \) denoting the number of mutants of genotype \( j \) evolving from the genotype-\( i \) subpopulation, and \( S_i \) denoting the size of the genotype-\( i \) subpopulation.

Since sum of Poisson random variables is a Poisson random variable, we can compute the distribution for the total number of emerging genotype-\( i \) mutants. Thus, the random variable

\[
    \eta_i = \sum_{j=1}^{g} \eta_{ij} \tag{3}
\]
is Poisson with mean
\[ \lambda_i = \mathbb{E}[\eta_i] = \sum_{j=1}^{g} (p_{ij} \times \mu \times S_i) = \mu \times S_i. \] (4)

Therefore, the probability that genotype-\(i\) subpopulation produces \(\eta_i = k_i\) mutants is
\[ \Pr(\eta_i = k_i) = \frac{\lambda_i^{k_i} e^{-\lambda_i}}{k_i!}. \] (5)

In our numerical simulations we sample a Poisson random variable with the distribution given by the equation above to generate the number of genotype-\(i\) mutants. Random variable \(\eta_i\) describes the total number of mutants produced by the genotype-\(i\) subpopulation. Next, these mutants need to “distributed” among all other genotypes with appropriate probabilities.

We consider only advantageous mutations so that cells of genotype \(i\) only mutate into stronger cells of genotype \(j\), with \(F_j > F_i\). Therefore, since genotypes are ordered by their fitness, the mutation matrix, \(P\) is upper triangular, i.e. \(p_{ij} = 0\) for \(j \leq i\), since mutations are advantageous and genotype \(i\) cannot mutate into itself. The adaptive behavior discussed above is of main interest in many practical applications, including the genetic evolution experiments of \(E.coli\). However, the approach described here can be easily extended to any mutation matrix, including the case with deleterious mutation, so that \(P\) is a full matrix.

As a particular example, we use the mutation matrix which describes one of the most common scenarios when every mutation occurrence in genotype-\(i\) population is equally likely to become any of the stronger mutants. In this case, the transition matrix is based on a uniform distribution among \((g - i)\) mutants and \(P\) becomes
\[
\mathcal{P} = \begin{pmatrix}
0 & \frac{1}{g-1} & \frac{1}{g-1} & \cdots & \frac{1}{g-1} \\
0 & 0 & \frac{1}{g-2} & \cdots & \frac{1}{g-2} \\
0 & \vdots & \ddots & \vdots & \vdots \\
0 & 0 & 0 & \cdots & 1 \\
0 & 0 & 0 & \cdots & 0
\end{pmatrix}
\] (6)

and entries of the matrix \(\mathcal{P}\) can be described as
\[ p_{ij} = \begin{cases} 
\frac{1}{g-i} & \text{if } i < j \\
0 & \text{otherwise}.
\end{cases} \] (7)

After the mutations occur sub-populations of each genotype become different and the conditional distribution for sub-populations after the mutation can be described using the multinomial distribution (corresponding to the nonzero \(p_{ij}\) entries) with
\[
\Pr(\eta_{i+1} = k_{i+1}, \eta_{i+2} = k_{i+2}, \ldots, \eta_g = k_g | \eta_i = k_i) = \frac{k_i!}{k_{i+1}!k_{i+2}!\cdots k_g!} p_{i+1}^{k_{i+1}} p_{i+2}^{k_{i+2}} \cdots p_g^{k_g}
\]
with
\[ \sum_{j=i+1}^{g} k_{ij} = k_i. \]
Consider a particular genotype-\(j\) subpopulation and define the number of emigrants and immigrants as
\[
Em_j = \sum_{i=1}^{g} \eta_{ij} \quad \text{and} \quad Im_j = \sum_{i=1}^{g} \eta_{ji},
\]
respectively. Then, the net change in the size of the genotype-\(j\) sub-population is
\[
\Delta S_j = Im_j - Em_j.
\]
Given particular values of \(\eta_{ij} = k_{ij}\) one can easily compute the changes in the size of genotype-\(j\) subpopulation due to mutations. In addition, one can show that mutation step does not alter the net population size, i.e. \(\sum_{j=1}^{g} \Delta S_j = 0\).

### 2.3 Dilution

During the dilution step, a random sample of size \(N\) is extracted (after growth and mutations) from the population of size \(S\). We model the dilution step using the multinomial distribution based on the frequencies \(f_j\) of the individual genotype subpopulations, with
\[
f_j = \frac{N_j \times F_j + \Delta S_j}{S}.
\]
Thus, letting \(\kappa_j\) denote the number of genotype-\(j\) cells selected after dilution, the distribution of \(\kappa_j, j = 1, \ldots, g\) follows
\[
Pr(\kappa_1 = k_1, \kappa_1 = k_2, \ldots, \kappa_g = k_g) = \left(\frac{N!}{k_1!k_2!\cdots k_g!}\right) f_1^{k_1} f_2^{k_2} \cdots f_g^{k_g}
\]
with
\[
\sum_{j=1}^{g} k_j = N.
\]

The process of growth, mutations, and dilution is repeated daily for the duration of a particular experimental run. The growth is deterministic, whereas the mutation and dilution steps are random. These three steps outline the adaptation model for a single isolated population. The process modeled by these three steps is a discrete-time Markov chain for the random variables \(H(t) = (H_1(t), \ldots, H_g(t))\) which describe the time-evolution of frequencies for the genotypes in the population. All frequencies sum up to one, i.e. \(\sum_{j=1}^{g} H_j(t) = 1\), and a sample path of this process can be easily generated given the initial distribution of genotypes \(H(1)\).

### 2.4 Coloring

A major element in this study is introducing a visually distinguishable feature of color that is inherited by all mutations. Without altering major properties of the cells, the entire population is partitioned into two subpopulations. It is colored with two biomarkers so that half of it becomes white and the other half, red. This use of neutral markers has been commonly used in the relevant studies (see e.g. [18, 13, 10, 5]) since this coloring does not alter the biology of the cells. The colors are inherited by the children and subsequent mutants emerging from the initial sub-population of a particular color.

The white and red coloring described above doubles the number of frequency variables in the model. There are now \(g\) white and correspondingly \(g\) red possible genotype subpopulations. The
model presented above for the single-color case can be easily extended to the two-color case by adding more elements to the frequency array $H$. Essentially, this is equivalent to introducing twice as many genotypes in the model. Thus, for the model with coloring, we introduce $H^w_j(t)$ and $H^r_j(t)$ for $j \in \{1, 2, \ldots, g\}$ representing the genotype-$j$ white sub-frequency and red sub-frequency, respectively. Altogether, we define $H^w(t) = \{H^w_1(t), \ldots, H^w_g(t)\}$ and $H^r(t) = \{H^r_1(t), \ldots, H^r_g(t)\}$ describing the frequency of $j$-genotype in the white and red sub-populations, respectively. The whole population is then described by the array of genotype frequencies $H(t) = \{H^w(t), H^r(t)\}$.

Please note that the growth and mutations steps are independent for the white- and red-sub-populations because the white and red cells cannot mutate into each other. Therefore, the growth and mutation steps for the white- and red-sub-populations can be analyzed and computed in parallel. However, the dilution step depends on the total size of the population and takes into account the frequency of occurrence of all white and red genotype-$j$ sub-populations in the whole population. Therefore, the formula (8) has to modified appropriately to describe the random selection of cells from $2g$ sub-populations, where $j = \{1, \ldots, g\}$ describes white genotype-$j$ sub-populations and $j = \{g+1, \ldots, 2g\}$ describes red genotype-$j$ sub-populations.

3 Estimation Method

An output of realistic experiments on genetic evolution is an ensemble of trajectories of different genotypes. In case of the *E. coli*, there is also a distinction between the white and red cells. This ensemble arises because the experiment is repeated many times and the initial conditions for each trajectory can be controlled very well and correspond to 50-50 proportion of white and red cells of genotype 1 (lowest fitness). For each trajectory, several mutation events can occur which corresponds to radical shifts in the relative proportion of white and red cells. However, for each particular trajectory, the precise time of mutations (i.e. which day) and the precise type of mutations is unknown. The experimental observer can only record changes in the relative proportion of white and red cells; this information is recorded daily with quite high accuracy. Therefore, our goal is to develop an optimization framework which will allow to accurately estimate (i) time of the mutation, (ii) fitness of the new emergent white- (or red-) cell mutant.

To achieve the goal outlined above, we concentrate here on trajectories which have only one or two mutations. This is motivated by biological experiments since each trajectory is recorded for a relatively short period of time (approximately 40 days), and the majority of trajectories only have one or two mutations. Trajectories with three or more mutations correspond to rare events and, in addition, such trajectories cannot be estimated accurately due to the relatively high sensitivity of later mutations (third, fourth, etc.) to small perturbations of parameters for earlier first and second mutation events.

To develop our estimation framework, we first derive explicit analytical expressions for the most-likely evolutionary trajectory conditioned on the *observed* composition of the bacterial population on the previous day, i.e. we assume that frequencies of bacterial genotype are known on the previous day, and we derive the most-likely population frequencies on the next day.

Some mutation events can be *unobserved* due to the fact that they are lost during the dilution step. Therefore, we develop explicit expressions for the most likely trajectory of the stochastic model outlined in section 2 given that there is only one *observed* mutation during the time-span of an evolutionary trajectory (however, the time of the mutation and the fitness of the emergent mutation are unknown) and we also develop expressions for the most-likely trajectory of the “locked-box” stochastic model given that there are only two *observed* mutations. We then develop an optimization procedure to compute numerically the best match for the time(s) of the mutation(s)
and selective advantage of the emergent genotype(s) to best match the observed white-to-red ratio in the overall *E. coli* population during the whole time-span of the evolutionary trajectory. Finally, since an ensemble of such trajectories is available, we are able to estimate the number of possible genotypes for each sub-population and their fitnesses.

### 3.1 Most-likely trajectory given that there is only one mutation

In this section we develop analytical expressions for the most-likely trajectory of the stochastic model outlined in section 2 given that there is only one observed mutation. In particular, we develop iterative formulas that yield expected frequencies based on frequency data on the previous day. The one-mutation trajectory has three unknown parameters - (i) day of the mutation, $\tau$, (ii) fitness of the emergent mutants, $F_\alpha$, and (iii) number of mutants $\eta$ that have emerged on day $\tau$. The optimization algorithm described later will use fitting techniques to optimize these unknown parameters of the most-likely trajectory for a particular observed frequency curve.

From here on, assume that there is only one observed mutation event which occurs at an unknown time $\tau$. At the beginning of the day $\tau$ and on each preceding day, there is only the one ancestor of genotype 1 in the whole population and no mutants, so that white and red frequencies $w_j(t) = r_j(t) = 0$, $\forall j \in \{2, \ldots, g\}$, $\forall t \leq \tau$. Note that the observed frequencies of the white- and red-cells, $w_1(t)$ and $r_1(t)$, respectively, are not necessarily equal to 1/2 for $t \leq \tau$, because of the stochasticity of the process. We will assume that there is a mutation in the white sub-population. Formulas for the mutation in the red sub-population are identical up to the index change.

At the beginning of day $(\tau + 1)$:

- There is an unknown number of mutants $\eta$ that have emerged in the white population after the daily selection (dilution) done at the end of day $\tau$. Denoting the (unknown) mutation rate by $\mu$, $\eta$ has a Poisson distribution with mean $\lambda = (1/2)N\mu F_1$ Assume that these $\eta$ emerged mutants are white and of unknown genotype $\alpha$ having an unknown growth factor $F_\alpha$.

At the beginning of day $(\tau + 1)$:

\[
\begin{align*}
    w_1(\tau + 1) &= w_1(\tau) - \eta/N, \\
    w_\alpha(\tau + 1) &= \eta/N, \\
    w_j(\tau + 1) &= 0, \ j \neq 1, \alpha, \\
    r_1(\tau + 1) &= r_1(\tau) = 1 - w_1(\tau), \\
    r_j(\tau + 1) &= 0, \ j = 2, \ldots, g.
\end{align*}
\]

After the mutation event, the process only exhibits growth and we can proceed iteratively for $t = (\tau + 1), (\tau + 2), \ldots, (k_f - 1), k_f$, where $k_f$ is the final day of the observations.

At beginning of day $t > \tau$:

The initial population contains only white cells of genotypes $j = 1, \alpha$ and only red cells of genotype $j = 1$, so that

\[
\begin{align*}
    w(t) &= w_1(t) + w_\alpha(t), \\
    w_1(t) &\geq 0, \ w_\alpha(t) \geq 0, \ r_1(t) \geq 0, \\
    r_j(t) &= 0, \ j = 2, \ldots, g, \\
    w_j(t) &= 0, \ j \neq 1, \alpha, \\
    r(t) &= r_1(t) = 1 - w(t).
\end{align*}
\]

Growth on day $t > \tau$:
Subpopulation sizes become:

- **white type** $1 \rightarrow \text{size} = NF_1w(t)$,
- **white type** $\alpha \rightarrow \text{size} = NF_\alpha w_\alpha(t)$,
- **red type** $1 \rightarrow \text{size} = NF_1r_1(t)$,

and all other sub-population sizes are zero. Introducing the variable $K(t)$ as

$$K \equiv K(t) = F_1w_1(t) + F_\alpha w_\alpha(t) + F_1r_1(t),$$

the total population size at the end of the growth phase of day $t$ becomes $NK$.

**Mutation on day $t > \tau$:**

On any given day $t > \tau$ various mutations can occur in both white- and red-subpopulations. However, these additional mutants are to be eliminated form the population during the dilution step. Therefore, denote the random numbers of mutants generated at the end of day $t$ as:

- number of mutants from type 1 whites $\rightarrow$ $j$ whites $= Nm_w^{1j}$,
- number of mutants from type $\alpha$ whites $\rightarrow$ $j$ whites $= Nm_w^{\alpha j}$,
- number of mutants from type 1 reds $\rightarrow$ $j$ reds $= Nm_r^{1j}$,

where $m_w^{ij}$ and $m_r^{ij}$ denote frequencies of genotypes after mutations in white and red sub-populations, respectively. Please note that the number of mutants ($Nm_w^{1j}$, $Nm_w^{\alpha j}$, and $Nm_r^{1j}$) defined above are Poisson random variables as defined in (2). For example, the conditional mean for the number of mutants 1-whites $\rightarrow$ $j$-whites is $E[Nm_w^{1j}|w_1(t)] = \tilde{\lambda} = \mu p_1 NF_1w_1(t)$, which defines $m_w^{1j}$ as Poisson random variables over rational numbers, i.e.

$$Pr\{m_w^{1j} = k/N|w_1(t)\} = \frac{\tilde{\lambda}^k e^{-\tilde{\lambda}}}{k!}$$

and conditional mean and variance of $m_w^{1j}$ are $E[m_w^{1j}|w_1(t)] = \mu p_1 NF_1w_1(t)$ and $Var\{m_w^{1j}|w_1(t)\} = \mu N^{-1} p_1 F_1w_1(t)$. Similar expressions hold for other two genotype frequencies after mutation.

**Dilution on day $t > \tau$:**

Conditioning on the following non-zero frequencies at the beginning of day $t$:

$$freq(t) = \{w_1(t), w_\alpha(t), r_1(t)\}$$

the dilution step must realize the following event $\Omega = \{\text{random sample size } N \text{ picked from terminal population of day } t \text{ contains only types 1 and } \alpha \text{ whites, and type 1 reds}\}$.

The conditional distribution of such random samples given that event $\Omega$ holds can be shown (using the multinomial properties and that the probabilities ignoring those frequencies are equal to 0) to be a multinomial $M(RestPop, N)$, where one extracts a random sample of size $N$ from the restricted population, defined by $RestPop = \{\text{subpopulation consisting of only the white-1, white-}\alpha, \text{ and red-1 cells existing at the terminal stage of day } t \text{ just before dilution}\}$.

The size of the population $RestPop$ is given by:

$$S = NK + N \left[ m_w^{\alpha 1} - \sum_{j=2}^{g} m_w^{1j} \right] + N \left[ m_{1\alpha}^{w} - \sum_{j=1,j\neq \alpha}^{g} m_w^{\alpha j} \right] - N \sum_{j=2}^{g} m_r^{1j} = N[K - H],$$

11
where $H$ is a Poissonian random variable which corresponds to the random fluctuations of the population due to mutations

$$H = \sum_{j=2}^{g} m_{1j}^w - m_{11}^w + \sum_{j=1,j\neq\alpha}^{g} m_{\alpha j}^w - m_{1\alpha}^w + \sum_{j=2}^{g} m_{1j}^r. \quad (12)$$

In particular, $H$ describes how many cells has left the population $RestPop$ due to mutations. Moreover, if we define random variables which correspond to changes in each sub-population as

$$H_1^w = \sum_{j=2}^{g} m_{1j}^w - m_{11}^w, \quad H_\alpha^w = m_{1\alpha}^w - \sum_{j=1,j\neq\alpha}^{g} m_{\alpha j}^w, \quad H_1^r = \sum_{j=2}^{g} m_{1j}^r, \quad (13)$$

then $H = H_1^w - H_\alpha^w + H_1^r$ and within the sub-population $RestPop$ the frequencies before the dilution of white-1, white-\alpha, and red-1 genotype are given by

$$q_1^w = \frac{F_1 w_1(t) - H_1^w}{K - H}, \quad q_\alpha^w = \frac{F_\alpha w_\alpha(t) + H_\alpha^w}{K - H}, \quad q_1^r = \frac{F_1 r_1(t) - H_1^r}{K - H}. \quad (14)$$

Note, that $H_1^w$, $H_\alpha^w$, and $H_1^r$ are $O(\mu \times \max\{F_i\})$, where $\mu$ is the mutation rate and $F_i$ are growth factors. In particular, we can compute the conditional means and variances of random variables in $\|13\|$. The conditional means are

$$\mathbb{E}[H_1^w|freq(t)] = \mu F_1 w_1(t) - \mu F_\alpha p_\alpha w_\alpha(t),$$

$$\mathbb{E}[H_\alpha^w|freq(t)] = \mu F_\alpha p_\alpha w_1(t) - \mu F_\alpha w_\alpha(t),$$

$$\mathbb{E}[H_1^r|freq(t)] = \mu F_1 r_1(t),$$

where we used $\sum_j p_{ij} = 1$. The conditional variance of $H$, given $freq(t)$, can be computed easily since mutation events are independent and conditional variances of $m_{i,j}^{w,r}$ can be computed explicitly. Therefore,

$$Var\{H|freq(t)\} = \sum Var\{m_{i,j}^{w,r}|freq(t)\} \leq 3\mu N^{-1} \max_j F_j, \quad (16)$$

where we used $\sum_j p_{ij} = 1$ and the fact that $w_1(t), w_\alpha(t), r_1(t) \leq 1$. Cases for $H_1^w$, $H_\alpha^w$, and $H_1^r$ are similar.

Therefore, frequencies in $\|14\|$ can be approximated by their first-order expansions

$$K q_1^w \approx F_1 w_1(t) - H_1^w + (H/K) F_1 w_1(t),$$

$$K q_\alpha^w \approx F_\alpha w_\alpha(t) + H_\alpha^w + (H/K) F_\alpha w_\alpha(t),$$

$$K q_1^r \approx F_1 r_1(t) - H_1^r + (H/K) F_1 r_1(t) \quad (17)$$

and the neglected terms in the expansion above are much smaller than the leading order terms due to the small mutation rates.

Please note that in the discussion above $K \equiv K(t)$, $H \equiv H(t)$, $H_1^w \equiv H_1^w(t)$, $H_\alpha^w \equiv H_\alpha^w(t)$, and $H_1^r \equiv H_1^r(t)$ are defined in terms of frequencies and mutants on day $t$. Therefore, $K$ can be treated as a constant in conditional expectation given $freq(t)$, and expected values of mutation terms on day $t + 1$ can be computed explicitly in terms of frequency information on day $t$.

Expected frequencies on day $t + 1$ with $t \geq 7$:

Given $freq(t)$ and that event $\Omega$ holds, the mean conditional frequencies after selection are identical to the existing frequencies before selection, i.e. $\mathbb{E}[w_j(t+1)|freq(t)] = \mathbb{E}[q_j^w|freq(t)]$ and $\mathbb{E}[r_j(t+1)|freq(t)] = \mathbb{E}[q_j^r|freq(t)]$ for $1 \leq j \leq g$. Therefore, if we define

$$W_j(t + 1) = \mathbb{E}[w_j(t + 1)|freq(t)] \quad \text{and} \quad R_j(t + 1) = \mathbb{E}[r_j(t + 1)|freq(t)]$$

with
then

\[
KW_1(t + 1) = K \mathbb{E}[q_w^1|f_{\text{req}}(t)] = \mathbb{E}[K q_w^1|f_{\text{req}}(t)]
\approx \mathbb{E}[F_1 w_1(t) - H_w^1 + (H/K) F_1 w_1(t)|f_{\text{req}}(t)]
= F_1 w_1(t) - \mathbb{E}[H_w^1|f_{\text{req}}(t)] + K^{-1} F_1 w_1(t) \mathbb{E}[|H|f_{\text{req}}(t)]
= (1 - \mu) F_1 w_1(t) + \mu F_\alpha p_{\alpha 1} w_\alpha(t) + \mu/K F_1 w_1(t) \left[(1 - p_{1\alpha}) F_1 w_1(t) + (1 - p_{\alpha 1}) F_\alpha w_\alpha(t) + F_1 r_1(t)\right]
= (1 - \mu) F_1 w_1(t) + \mu F_\alpha p_{\alpha 1} (w(t) - w_1(t)) + \mu/K F_1 w_1(t) \times \left[(1 - p_{1\alpha}) F_1 w_1(t) + (1 - p_{\alpha 1}) F_\alpha (w(t) - w_1(t)) + F_1 (1 - w(t))\right],
\]

where we’ve taken into account that \(w_\alpha(t) = w(t) - w_1(t)\). Similarly, \(K(t)\) can be expressed as

\[
K(t) = F_1 + (F_\alpha - F_1)(w(t) - w_1(t)),
\]

where we also used that \(r_1(t) = 1 - w(t)\).

Note, that \(w_1(t)\) is the only unknown parameter in equations (18) and (19) since the frequency of the type-1 and type-\(\alpha\) white cells can only be observed together, i.e. \(w(t) = w_1(t) + w_\alpha(t)\). However, we can approximate the unknown frequency of type-1 white cells on day \(t\) by its expectation, and arrive at the approximation of \(KW_1(t + 1)\)

\[
KW_1(t + 1) \approx KKW_1(t + 1) = (1 - \mu) F_1 W_1(t) + \mu F_\alpha p_{\alpha 1} (w(t) - W_1(t)) + \mu/K F_1 W_1(t) \left[(1 - p_{1\alpha}) F_1 W_1(t) + (1 - p_{\alpha 1}) F_\alpha (w(t) - W_1(t)) + F_1 (1 - w(t))\right],
\]

where \(w(t)\) is the observed frequency of white cells on day \(t\), and \(W_1(t)\) is the expected frequency of type-1 white cells computed form the previous iteration, and we approximated \(K(t)\) as

\[
K(t) \approx K(t) = F_1 + (F_\alpha - F_1)(w(t) - W_1(t)).
\]

Similarly, we can develop an approximation for \(KW_\alpha(t + 1)\)

\[
KW_\alpha(t + 1) \approx KKW_\alpha(t + 1) = (1 - \mu) F_\alpha (w(t) - W_1(t)) + \mu F_1 p_{1\alpha} W_1(t) + \mu/K F_\alpha (w(t) - W_1(t)) [F_1 (1 - p_{1\alpha})(1 - w(t)) + F_\alpha (1 - p_{\alpha 1})(w(t) - W_1(t))].
\]

Please note, that the expressions above are general and do not make an assumption that the mutations are not reversible (\(p_{\alpha 1} = 0\)).

### 3.2 Most-likely trajectory given that there two events

In this section we develop formulas for the most-likely trajectory with two mutation events. To follow the notation in the previous section, we assume that the first event is described by the triple \(\{\tau, \alpha, \eta\}\), where \(\tau\) is the time of the mutation, \(\alpha\) is the genotype after the mutation, \(\eta\) is the number of mutants. Without the loss of generality, we can assume that the first mutation event occurs in the white sub-population (if this is not the case, just rename the colors). However, the second event can occur in the same sub-population (white) or in the opposite sub-population (red). We will develop expressions for the most-likely trajectory for both cases.

The second mutation event is described by the parameters \(\{\tau_2, \beta, \eta_2\}\), where \(\tau_2 > \tau\) is the time of the second mutation, \(\beta\) is the mutation genotype, and \(\eta_2\) is the number of mutants. To
detect significant second events, the genotype in the second mutation should be stronger that the
genotype of the first mutation, otherwise, the second event is not detectable with a significant level
of confidence. Therefore, we assume that $\beta > \alpha > 1$.

The analysis of the two-event trajectory is quite similar to the analysis presented in section 3.1
and we present here only the final expressions for the most-likely trajectory with two events.

3.2.1 Same Colors

Here we present formulas for the most-likely trajectory with two mutations in the white sub-
population. Then, on day $t+1$ (for $t > \tau_2$), conditioned on frequencies $freq(t) = \{w_1(t), w_\alpha(t), w_\beta(t), r_1(t)\}$
(all other frequencies are zero) we obtain

\[
KW_1(t + 1) \approx F_1w_1(t) - \mu(F_1w_1(t) - p_{\alpha_1}F_\alpha w_\alpha(t) - p_{\beta_1}F_\beta w_\beta(t)) + (\mu/K)F_1w_1(t)\times
\]
\[
(1 - p_{\alpha_1} - p_{\beta_1})F_1w_1(t) + (1 - p_{\alpha_1})F_\alpha w_\alpha(t) + (1 - p_{\beta_1})F_\beta w_\beta(t) + F_1r_1(t),
\]  

(23)

\[
KW_\alpha(t + 1) \approx F_\alpha w_\alpha(t) - \mu(F_\alpha w_\alpha(t) - p_{\alpha_1}F_1w_1(t) - p_{\beta_1}F_\beta w_\beta(t)) + (\mu/K)F_\alpha w_\alpha(t)\times
\]
\[
(1 - p_{\alpha_1} - p_{\beta_1})F_1w_1(t) + (1 - p_{\alpha_1})F_\alpha w_\alpha(t) + (1 - p_{\beta_1})F_\beta w_\beta(t) + F_1r_1(t),
\]  

(24)

\[
KW_\beta(t + 1) \approx F_\beta w_\beta(t) - \mu(F_\beta w_\beta(t) - p_{\alpha_1}F_1w_1(t) - p_{\beta_1}F_\alpha w_\alpha(t)) + (\mu/K)F_\beta w_\beta(t)\times
\]
\[
(1 - p_{\alpha_1} - p_{\beta_1})F_1w_1(t) + (1 - p_{\alpha_1})F_\alpha w_\alpha(t) + (1 - p_{\beta_1})F_\beta w_\beta(t) + F_1r_1(t),
\]  

(25)

where $K(t) = F_1w_1(t) + F_\alpha w_\alpha(t) + F_\beta w_\beta(t) + F_1r_1(t)$. Since $r_1(t) = 1 - w(t)$ and $w_\beta(t) =
w(t) - w_1(t) - w_\alpha(t)$, $K(t)$ becomes

\[
K(t) = F_1 + (F_\beta - F_1)(w(t) - w_1(t)) - (F_\beta - F_\alpha)w_\alpha(t).
\]  

(26)

Therefore, we can compute formulas in (23), (24), (25) compute iteratively by approximating
$w_1(t) \approx W_1(t)$, $w_\alpha(t) \approx W_\alpha(t)$ and substituting into the expressions above.

Different Colors

Here we develop formulas for the most likely trajectory when the second mutation event occurs in
the red sub-population. Then, on day $t+1$ conditioned on frequencies $freq(t) = \{w_1(t), w_\alpha(t), r_1(t), r_\beta(t)\}$
we obtain

\[
KW_1(t + 1) \approx F_1w_1(t) - \mu(F_1w_1(t) - p_{\alpha_1}F_\alpha w_\alpha(t)) + (\mu/K)F_1w_1(t)\times
\]
\[
(1 - p_{\alpha_1})F_1w_1(t) + (1 - p_{\alpha_1})F_\alpha w_\alpha(t) + (1 - p_{\beta_1})F_1r_1(t) + (1 - p_{\beta_1})F_\beta r_\beta(t),
\]  

(27)

\[
KW_\alpha(t + 1) \approx F_\alpha w_\alpha(t) + \mu(p_{\alpha_1}F_1w_1(t) - F_\alpha w_\alpha(t)) + (\mu/K)F_\alpha w_\alpha(t)\times
\]
\[
(1 - p_{\alpha_1})F_1w_1(t) + (1 - p_{\alpha_1})F_\alpha w_\alpha(t) + (1 - p_{\beta_1})F_1r_1(t) + (1 - p_{\beta_1})F_\beta r_\beta(t),
\]  

(28)

\[
KR_1(t + 1) \approx F_1r_1(t) - \mu(F_1r_1(t) - p_{\beta_1}F_\beta r_\beta(t)) + (\mu/K)F_1r_1(t)\times
\]
\[
(1 - p_{\alpha_1})F_1w_1(t) + (1 - p_{\alpha_1})F_\alpha w_\alpha(t) + (1 - p_{\beta_1})F_1r_1(t) + (1 - p_{\beta_1})F_\beta r_\beta(t),
\]  

(29)

\[
KR_\beta(t + 1) \approx F_\beta r_\beta(t) + \mu(p_{\beta_1}F_1r_1(t) - F_\beta r_\beta(t)) + (\mu/K)F_\beta r_\beta(t)\times
\]
\[
(1 - p_{\alpha_1})F_1w_1(t) + (1 - p_{\alpha_1})F_\alpha w_\alpha(t) + (1 - p_{\beta_1})F_1r_1(t) + (1 - p_{\beta_1})F_\beta r_\beta(t),
\]  

(30)
where $K(t) = F_1 w_1(t) + F_\alpha w_\alpha(t) + F_1 r_1(t) + F_\beta r_\beta(t)$. Since $w_\alpha(t) = w(t) - w_1(t)$ and $r_\beta(t) = r(t) - r_1(t) = (1 - w(t)) - r_1(t)$, expression for $K(t)$ becomes

$$K(t) = F_\beta - (F_\alpha - F_1) w_1(t) - (F_\beta - F_1) r_1(t) - (F_\beta - F_\alpha) w(t).$$

(31)

Similar to the previous sections, we approximate $w_1(t) \approx W_1(t)$ and $r_1(t) \approx R_1(t)$ to compute the expressions above interatively.

4 Simulated Experiments

Given the expressions in (19), (20), and (22), we can formulate the optimization problem to estimate parameters of the underlying population evolution stochastic model. In particular, we assume that the mutation rate, $\mu$, the mutation matrix $p_{ij}$, and the fitness of the ancestor genotype, $F_1$, are known. The goal of our approach is to estimate the selective advantage $s_\alpha$ given by (1). In addition, parameters $\tau$ (the time of mutation) and $\eta$ (number of emerging mutants) are also unknown and should be optimized.

We use the simulated data to illustrate our approach. We use the following parameters in our simulations and consequent reconstruction of the model’s fitnesses -

- $g = 4$, the number of allowed genotypes,
- $n = 1000$, the number of simulated trajectories,
- $N = 50000$, initial population size,
- $\vec{s} = [s_1, s_2, \ldots, s_g] = [0, 0.04, 0.07, 0.13]$, the selective advantages.

We considered the mutation matrix

$$P = \{p_{ij}\} = \begin{pmatrix}
0 & 1 & 1 & 1 \\
0 & 0 & 1 & 1 \\
0 & 0 & 0 & 1 \\
0 & 0 & 0 & 0 
\end{pmatrix}$$

with mutation rates $\mu = 10^{-6}$. These parameters are consistent with numerical values reported in past and ongoing experiments (see e.g. [9]).

After simulating the $n = 1000$ experiments (trajectories), day-to-day frequency data was stored. It consisted of the white and red population frequencies along with individual mutant frequencies of each color, at the days’ beginnings. Examples of trajectories with one mutation and two mutations are presented in Figures 1 and 2, respectively. We considered that a trajectory reached a fixation if $w(T_{th}) \geq 95\%$ (or $\leq 5\%$) and stopped the simulation at $t = T_{th}$. We allowed the maximum length of the trajectory in the absence of fixation to be $T_{final} = 200$ days. We also considered that a mutation event has occurred when a mutant frequency reached 0.1\% of the population size. Approximately 70\% and 23\% of trajectories had 1 and 2 mutation event, respectively. Approximately 0.5\% of trajectories had no mutation events and the average number of events before fixation was approximately 1.34.

We would like to emphasize that in experimental data only the frequency of the overall white population (solid line $w_{tot}$) can be observed. The overall frequency of the red sub-population can also be observed, but does not carry any additional information. Therefore, the goal of this work is determine the time of emergence of white or red mutants only from the overall frequency of the white sub-population.
Figure 1: Examples of two trajectories with one mutation showing the total frequency of white-cell population and a mutant sub-population. Black line (labeled \( w_{\text{tot}} \)) denotes the total size of the white-cell population. Left part - emergence of the red genotype-2 mutant (red line labeled \( r_2 \)), Right part - emergence of the white genotype-4 mutant (blue line labeled \( w_4 \)).

Figure 2: Examples of two trajectories with two mutations showing the total frequency of white-cell population and mutant sub-populations. Black line (labeled \( w_{\text{tot}} \)) denotes the total size of the white-cell population. Left part - emergence of the red genotype-3 mutant (red line labeled \( r_3 \)) and later emergence of the white genotype-2 mutant (blue line labeled \( w_2 \)), Right part - emergence of the red genotype-2 mutant (red line labeled \( w_2 \)) and later emergence of white genotype-4 mutant (blue line labeled \( w_4 \)).

5 Fitness Estimation

In this section we formulate an optimization problem for computing the selective advantage of the emergent genotype. We will also develop an automatic test to verify whether each trajectory has one or two mutation events.

5.1 First Mutation Event

We will first apply expressions (20) and (22) for one mutation occurrence to define the optimization problem given that only one mutation occurred. We would like to point out that these expressions depend on four parameters - (i) the time of emergence, \( \tau \), (ii) number of mutants emerging on the day \( \tau + 1 \), \( \eta \), (iii) the genotype \( \alpha \), and (iv) selective advantage of the mutant, \( s_\alpha \). The choice of \( \alpha \) in (32) impacts which entries of the mutation matrix \( p_{\beta \alpha} \) and \( p_{\alpha \beta} \) are used in computing \( W_1(t + 1) + W_\alpha(t + 1) \) in (20) and (22), respectively. We would like to note that the expected frequency of the white cells during the mutation can be computed as \( W_1(\tau) = w(\tau) \), \( W_1(\tau + 1) = w(\tau) - \eta/N \),
\[ W_\alpha(\tau) = 0, W_\alpha(\tau + 1) = \eta/N. \] For \( t > \tau + 1 \) they can be computed using recursive formulas (20) and (22).

Therefore, we define the following first-event optimization problem

\[ s_\alpha = \arg \min_{\tau, \eta, \alpha, s_\alpha} \frac{1}{(t_f - t_i)} \sum_{t=t_i}^{t=t_f} (W(t) - w(t))^2, \quad (32) \]

where \( W(t) = W_1(t) + W_\alpha(t) \). We select the initial and final times for optimization as \( t_i = \tau \) and \( t_f = \min\{t \text{ s.t. } w(t) = th\} \), where \( th \) is the threshold value to determine when the genotype \( \alpha \) sufficiently developed in the population. Typical values we used are \( th = 0.55, \ldots, 0.65 \). We use the value of the threshold \( th < 0.7 \) in order to use the rest of the trajectory to define the optimization problem for the second-mutation event.

It is possible to implement a number of strategies for adjusting the final time of estimation, \( t_f \). In particular, one can use varying \( t_f \) and estimate the goodness of fit using the cost function in the right-hand side of (32). It is also possible to include \( t_f \) into the optimization problem in (32), i.e. optimize the cost function with respect to the final time \( t_f \). However, we expect that more refined procedures for determining \( t_f \) would produce similar results, since the optimization procedure in (32) already produces a fairly narrow histograms of possible selective advantages, \( s_\alpha \), as demonstrated later in this section.

\subsection*{5.2 Second Mutation Event}

After the optimization problem for the first mutation event has been computed, we can use the rest of the trajectory data (i.e. \( w(t) \) with \( t \in [t_f, T_{\text{max}}] \)) to define an optimization problem for the second mutation event. However, first, we need define an explicit criteria for the occurrence of the second mutation event. To this end, we define that a second mutation occurred if there is a time \( t^* > \tau \) such that

\[ |W_1(t^*) - w(t^*)| > th_2 \]

where \( th_2 = 0.02 \) is the threshold value for the second event. We assume that the color of the mutant is red is \( W_1(t^*) - w(t^*) > 0 \), and white otherwise. If there is no time \( t^* \) which satisfies the above criteria, we assume that the trajectory does not contain a second mutation event. We would like to remind, that we can assume that the first mutation event always occurs in the white sub-population (if this is not the case, just rename the colors). Therefore, if the second event occurs in the white sub-population, we can use formulas outlined in section 3.2.1 otherwise we use formulas for mutation events in sub-populations of different colors in section 3.2.1. Thus, we can define an optimization procedure similar to (32) using the second-event data. The optimization problem for the second-event data would yield a quadruple of values \( (\tau_2, \eta_2, \beta, s_\beta) \) where \( \beta \) is the genotype after mutation and \( s_\beta \) is its selective advantage. Therefore, the second-event data contains additional information compared to the first-event calculations, and this information can be used to refine the data computed from the optimization procedure for the first-event data.

\subsection*{5.3 Multi-Gaussian Fitting to Histogram of Selected Advantages}

Optimization procedures defined in section 5.1 and 5.2 produce one or two numbers which correspond to estimation of the selective advantages in mutated genotypes for a single observed trajectory. However, in typical evolutionary experiments many bacterial colonies are allowed to grow in parallel and, thus, there are many trajectories available for fitness estimation. Due to numerical
errors and various assumptions when deriving the optimization formulas, the optimization procedures in 5.1 and 5.2 are extremely unlikely to produce exactly the same values for the estimation of selective advantages for different trajectories.

Thus, when applied to multi-trajectory data, the optimization procedures in 5.1 and 5.2 yields a histogram of possible values for selective advantages. This histogram may have several peaks, but to compute the final estimates for the selective advantages a multi-Gaussian fit to this histogram is applied. In particular, we fit the function of the form

$$G(x) = \sum_{i=1}^{g-1} a_i e^{-\frac{x-b_i}{c_i^2}}$$

(33)

to the histogram data. The fitting is based on the sum of squared errors minimization and results in the \((g-1)\)-mode Gaussian \(\hat{G}(x)\). The centers of the multi-Gaussian fit, \(\hat{b}_i \equiv \hat{s}_i\) are then selected as the estimates of selective advantages in the overall population. The variances, \(\hat{c}_i\) also contain important information which can be interpreted as error bounds for the corresponding estimate.

6 Numerical Results

We illustrate our approach on the simulated data for \(n = 1000\) trajectories. The parameter values have been defined in section 4. It is very likely (\(> 99\%\)) that a first mutation event occurs in a trajectory for the mutation rate chosen in these experiments. However, multiple event occurrences become less common (\(\approx 37\%\) of trajectories). Moreover, if there are multiple events, they often occur close to each other time time, which presents a challenge of estimating their event times and using the second-event data efficiently and reliably.

Optimization procedure in (32) is carried out by a straightforward search through lists of possible values for \(\tau\), \(\eta\), and \(s_\alpha\). The following lists were used in computations

- \(LIST_{s_\alpha} = \{0.01, 0.02, 0.03, \ldots 0.20\}\)
- \(LIST_\tau = \{T_{min}, \ldots T_{max}\}\)
- \(LIST_\eta = \{50, 100, 150, \ldots 2000\}\)

with \(T_{max}\) chosen as the time at which there was at least a 5\% rise (or fall) of the white frequency, \(w(t)\), from .5. \(T_{min}\) was chosen as the maximum time prior to \(T_{max}\) such that there was at least a 0.5\% rise (or fall) from .5. The threshold values 5\% and 0.5\% were typical values used in all computations; these values did not significantly affect the results of the optimization procedure.

Figure 3 depicts the multi-Gaussian fit for the histogram of selective advantages obtained using the optimization procedure with the first-mutation data from \(n = 1000\) simulated trajectories. Selective advantages are estimated quite accurately for genotypes \(g = 2, 3, 4\). In particular, the histogram has distinct three local maxima, which is an indication that the developed estimation algorithm correctly captures the first-event data.

To demonstrate the applicability of the estimation procedure for the second-mutation data, we generate \(n = 10000\) trajectories. Approximately 3700 trajectories have the second-mutation events with 3280 different-color first and second mutation cases, and 420 same-color cases. Figure 4 depicts estimation of selective advantages from the first-mutation data (top part, denote the histogram as \(H_1\)), second-mutation data (middle part, denote the histogram as \(H_2\)), and the combination of the two events (bottom part, histogram \(H_{1,2}\)).
Figure 3: Multi-Gaussian fit for the estimation of genotype fitnesses from the optimization procedures for the first-mutation event from $n=1000$ simulated trajectories. The empirical histogram depicts the numbers of trajectories (runs) which produce the corresponding selective advantage; dashed line - function $G(x)$. The corresponding variances are also presented in the plot.

There is much fewer data for the second-mutation event, compared to the first-mutation event, which results in much larger variances for the multi-Gaussian fit. We also established empirically that the same-color mutation events are quite challenging for the correct estimation of the selective advantage for the second mutation. Therefore, it might be beneficial to remove same-color mutation data from the histogram of the second-mutation event.

Histogram $H_1$ for the estimation from the first-mutation event (top of Figure 4), and the combined histogram $H_{1,2}$ (bottom of Figure 4) are quite similar and the variances even increase slightly for the multi-Gaussian fit for $H_{1,2}$. However, these results demonstrate the overall applicability of the two-event estimation procedure and its robustness. Estimator $\hat{s}_1$ even becomes more accurate in the multi-Gaussian fit for $H_{1,2}$. Therefore, second-event data can be used to verify the accuracy of the first-event estimation. Moreover, if there are mutations with different rates, then some mutations can potentially only appear as second events. In this case, selective advantages for mutations with smaller mutation rates can only be estimated from the second-event data.

7 Conclusions

In this paper we present a numerical approach to estimate selective advantages for mutant genotypes using the data of long-term bacterial evolutionary experiments. Initially, we assume that bacterial calls are colored with a particular biomarker which allows to record time-evolution of the size of bacterial colonies of two different colors in a bacterial population. An ensemble of such trajectories is then used to estimate the histogram of selective advantages of mutant genotypes. A multi-Gaussian fit is utilized to analyze this histogram and estimate selective advantages. We demonstrated the applicability of our approach on simulated data and now it stands ready to be applied to more realistic observational data. Additional simulations and parameter studies can be found in [24].

Realistic experiments include a rather small number of populations growing in parallel. However, these experiments are carried out for many years, and bacterial populations are often re-seeded with the original ancestor genotypes after a dominant mutation occurs and one color overtakes the whole population. Therefore, a realistic number of trajectories available for estimating the selective
advantages are $O(100)$. It has been demonstrated in [24] that the variances in the multi-Gaussian fit (33) scale as $n^{-1/2}$. Nevertheless, estimation procedure described here is still applicable even with such as small number of available trajectories.

The main advantage of the estimation procedure developed here is that second-mutation events are taken into account. The fitness estimation from the second-mutation data can be sensitive to the results of estimation using the first-mutation data. This can be used to analyze the sensitivity of the estimation and demonstrate robustness, i.e. if the second-event estimates alter the histogram is selective advantages in a significant way, then one should recompute the estimates from the first-mutation event data. Finally, selecting appropriate thresholds to determine the occurrence of the second mutation can also influence the estimation. We plan to address this issue in a subsequent paper on application of our estimation framework to realistic observational data.
Figure 4: Multi-Gaussian fit for the estimation of genotype selective advantages from $n = 10000$ simulated trajectories and the corresponding fit $\hat{G}(x)$. Top part - estimation of selective advantages from the first-mutation data. Middle part - estimation of selective advantages from the second-mutation data. Bottom part - estimation of selective advantages from combining the first and second mutation events.
References

[1] J. E. Barrick and R. E. Lenski. Genome dynamics during experimental evolution. *Nature Reviews. Genetics*, 14(12):827–839, 2013.

[2] J. E. Barrick, C. C. Streltsoff, R. E. Lenski, and M. R. Kauth. *Escherichia coli* rpoB mutants have increased evolvability in proportion to their fitness defects. *Molecular Biology and Evolution*, 27(6):1338–1347, 2010.

[3] V. S. Cooper, D. Schneider, M. Blot, and R. E. Lenski. Mechanisms causing rapid and parallel losses of ribose catabolism in evolving populations of *Escherichia coli*. *Journal of Bacteriology*, 183:2834–2841, 2001.

[4] J. A. Moura de Sousa, P. R. A. Campos, and I. Gordo. An abc method for estimating the rate and distribution of effects of beneficial mutations. *Genome Biology and Evolution*, 5(5):794–806, 2013.

[5] J. A. Moura de Sousa, P. R.A. Campos, and I. Gordo. An ABC method for estimating the rate and distribution of effects of beneficial mutations. *Genome Biology and Evolution*, 5(5):794–806, 2013.

[6] Daniel E. Deatherage, Kepner L. Jamie, Albert F. Bennett, Richard E. Lenski, and Jeffrey E. Barrick. Specificity of genome evolution in experimental populations of *Escherichia coli* evolved at different temperatures. *Proc. Nat. Acad. Sci. USA*, 114(10):E1904–E1912, 2017.

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

[8] J. W. Fox and R. E. Lenski. From here to eternity - the theory and practice of a really long experiment. *PLoS Biology*, 13(6):e1002185, 2015.

[9] I. Gordo, L. Perfeito, and A. Sousa. Fitness effects of mutations in bacteria. *Journal of Molecular Microbiology and Biotechnology*, 21(1-2):20–35, 2012.

[10] Matthew Hegreness, Noam Shoresh, Daniel Hartl, and Roy Kishony. An equivalence principle for the incorporation of favorable mutations in asexual populations. *Science*, 311(5767):1615–1617, 2006.

[11] C. Illinworth and V. Mustonen. A method to infer positive selection from marker dynamics in asexual population. *Bioinformatics*, 28(6):831–837, 2012.

[12] M. Kimura. On the probability of fixation of mutant genes in a population. *Genetics*, 47:713–719, 1962.

[13] R. Korona, C. H. Nakatsu, L. J. Forney, and R. E. Lenski. Evidence for multiple adaptive peaks from populations of bacteria evolving in a structured habitat. *Proc. Nat. Acad. Sci. USA*, 91(19):9037–9041, 1994.

[14] G. I. Lang, D. Botstein, and M. M. Desai. Genetic variation and the fate of beneficial mutations in asexual populations. *Genetics*, 188:647–661, 2011.

[15] R. E. Lenski, M. R. Rose, S. C. Simpson, and S. C. Tadler. Long-term experimental evolution in *Escherichia coli*. i. adaptation and divergence during 2000 generations. *The American Naturalist*, 138(6):1315–1341, 1991.
[16] F. Peng, S. Widmann, A. Wünsche, K. Duan, K. A. Donovan, R. C. J. Dobson, R. E. Lenski, and T. F. Cooper. Effects of beneficial mutations in pykF gene vary over time and across replicate populations in a long-term experiment with bacteria. *Molecular Biology and Evolution*, 10(6):msx279, 2017.

[17] L. Perfeito, L. Fernandes, C. Mota, and I. Gordo. Adaptive mutations in bacteria: high rate and small effects. *Science*, 317:813–815, 2007.

[18] L. D. Plank and J. D. Harvey. Generation time statistics of *Escherichia coli* B measured by synchronous culture techniques. *Journal of General Microbiology*, 115:69–77, 1979.

[19] S. H Rice. *Evolutionary Theory*. Sinauer Associates, 2004.

[20] D. R. Rokyta, C. J. Beisel, P. Joyce, M. T. Ferris, C. L. Burch, and H. A. Wichman. Beneficial fitness effects are not exponential for two viruses. *Journal of Molecular Evolution*, 67:368–376, 2008.

[21] D. R. Rokyta, P. Joyce, S. B. Caudle, and H. A. Wichman. An empirical test of the mutational landscape model of adaptation using a single-stranded DNA virus. *Nature Genetics*, 37:441–444, 2005.

[22] S. M. Ross. *Stochastic Processes*. John Wiley & Sons, New York, 1996.

[23] D. Rozen, J. de Visser, and P. Gerrish. Fitness effects of fixed beneficial mutations in microbial populations. *Current Biology*, 12:1040–1045, 2002.

[24] Sergey S. Sarkisov. *Parameters Estimation for Stochastic Genetic Evolution of Asexual Populations*. PhD thesis, University of Houston, 2017.

[25] D. Simon. *Evolutionary Optimization algorithms*. John Wiley & Sons, New York, 2013.

[26] F. Vasi, M. Travisano, and R. Lenski. Long-term experimental evolution in *Escherichia coli*. ii. changes in life-history traits during adaptation to a seasonal environment. *The American Naturalist*, 144(3):432–456, 1994.

[27] R. J. Woods, J. E. Barrick, T. F. Cooper, U. Shrestha, M. R. Kauth, and R. E. Lenski. Second-order selection for evolvability in a large *Escherichia coli* population. *Science*, 331(6023):1433–1436, 2011.

[28] W. Zhang, V. Sehgal, D. Dinh, R. R. Azevedo, T. Cooper, and R. Azencott. Estimation of the rate and effect of new beneficial mutations in asexual populations. *Theoretical Population Biology*, 81(2):168–178, 2012.