ESTIMATION OF MUTATION RATES AND SELECTIVE ADVANTAGES IN CELL POPULATION DYNAMICS

A Dissertation

Presented to

the Faculty of the Department of Mathematics

University of Houston

In Partial Fulfillment

of the Requirements for the Degree

Doctor of Philosophy

By

Vasudha Sehgal

December 2011

ESTIMATION OF MUTATION RATES AND SELECTIVE ADVANTAGES IN CELL POPULATION DYNAMICS

Vasudha Sehgal

APPROVED:

Dr. Robert Azencott, Chairman
Department of Mathematics, University of Houston

Dr. Krešimir Josić
Department of Mathematics, University of Houston

Dr. Timothy F. Cooper
Department of Biology and Biochemistry,
University of Houston

Dr. Ilya Timofeyev
Department of Mathematics, University of Houston

Dean, College of Natural Sciences and Mathematics

ESTIMATION OF MUTATION RATES AND SELECTIVE ADVANTAGES IN CELL POPULATION DYNAMICS

An Abstract of a Dissertation

Presented to

the Faculty of the Department of Mathematics

University of Houston

In Partial Fulfillment

of the Requirements for the Degree

Doctor of Philosophy

By

Vasudha Sehgal

December 2011

Abstract

Adaptation of a population to a new environment is driven by the emergence of beneficial mutations and their spread due to natural selection. For this reason, the rate and effect of the beneficial mutations are key parameters in determining the degree of adaptation of a population in a new environment. Stochastic models for this process have been developed, however, many of the relevant population parameters are poorly known, largely due to the difficulty of using experiments to understand the underlying stochastic process of mutations. In this thesis, we analyze the experiments that track the dynamics of neutral markers, for the evolving asexual populations of bacteria *Escherichia coli* to study the effect of newly arising beneficial mutations. We present a new simulation approach, to estimate the rate and size of these beneficial mutations, and to develop efficient estimators of mutation rates and selective advantages. We evaluate the accuracy of these estimators through our comprehensive simulations. These estimators are quite robust to the use of relatively low experimental replications. To study the validity of our model, we compare experimentally determined estimates of selective advantages to our theoretically obtained estimates of selective advantages. We find that our theoretical predictions are not very different from selective coefficients obtained experimentally. We perform the study first under the simplifying assumption that only one irreversible mutation is available to the population, and then extend this to a model that allows multiple mutations to be available to the population. Application of our method to suitably designed experiments will allow estimation of how evolvability of population depends on demographic and initial fitness parameters.

Contents

List of Figures

List of Tables

CHAPTER 1

Introduction

1.1 Historical Background

In the introduction to The Origin of Species, Darwin (1859) [9] states the following:

> "As many more individuals of each species are born than can possibly survive; and as, consequently, there is a frequently recurring struggle for existence, it follows that any being, if it vary however slightly in any manner profitable to itself, under the complex and sometimes varying conditions of life, will have a better chance of surviving and thus naturally

> selected. From the strong principle of inheritance, any selected variety will tend to propagate its new and modified form. "

Although much progress has been made in biology since Darwin's time, his theory of natural selection still remains as the only scientifically acceptable theory to explain why organisms are so well adapted to their environments. As an intuitive idea for the definition of adaptation, Rice (1961) [47] proposes that organisms evolve traits that maximize size of the population of those organisms in a particular environment. Adaptive evolution is driven by the emergence of beneficial changes in the DNA sequence of a cell's genome, and their subsequent spread due to natural selection. The occurrence of beneficial changes in a cell's genome is termed "Mutation". New factors arise in an organism by the process of mutation. If a mutant with high selective advantage appears in a population, then it has a positive probability of survival, (Kimura (1983) [32]) however large the population may be. Haldane (1927) [21] in his work about developing a mathematical theory for natural selection, shows that in a constant size population, the probability that a mutation with selective advantage s will survive random changes in allele frequency due to random sampling is approximately $2s$. These changes in allele frequency in a population due to random sampling are termed "Genetic Drift".

In population genetics, it is assumed that selection acts on individuals based on their phenotype and these phenotypes are determined by individuals' genotype. Thus, the distinction of fitness as a property of an individual or as a property of a genotype is not an issue in population genetics (Rice (1961) [47]). Fitness defines the ability of an individual to both survive and reproduce in an environment. The distribution of

fitness reflects the selection coefficient relative to the fittest alleles in the population.

It is common in the literature to assume that there is an underlying distribution of absolute fitnesses from which the relative fitnesses are derived. The absolute fitnesses are the expected number of surviving or successful offsprings. The relative fitnesses are simply the values of absolute fitnesses scaled in some way. For example, one could obtain relative fitnesses if one divides the fitness value of each genotype by the largest absolute fitness value so that the fittest genotype has a relative fitness of 1. There are two main fitness distributions that have been commonly used. Ohta (1977) [43] investigated a model in which selection coefficients against the mutants follow an exponential distribution; if σ denotes the standard deviation characterizing the dispersion in the distribution, and if s denotes the selective advantage, then the exponential distribution is given by

$$f_e(s) = \frac{1}{\sigma} e^{s/\sigma}, \quad s < 0.$$

Kimura (1979) [31] restricted the study to deleterious mutations, and disregarded beneficial mutations. He studied a model of nearly neutral mutations that assumes that the selection coefficient against mutant at various sites within a gene follows a gamma distribution given by

$$f(s) = \frac{\alpha^\beta}{\Gamma(\beta)} (-s)^{\beta-1} e^{\alpha s} \quad s < 0$$

where $\alpha = \beta/\sigma$. This model is based on the idea that selective neutrality is the limit when the selective advantage becomes indefinitely small. Kimura was led to the gamma distribution precisely, because, as he states in his (Kimura (1969) [31]) paper,

> "Ohta's model has a drawback in that it cannot accommodate enough mutations that behave effectively as neutral when the population size gets large. "

The gamma distribution has considerably more probability mass near zero, suggesting that selective coefficients would be much larger than for the exponential distribution. In 1991, Gillespie [19] gave an argument using extreme value theory, to conclude that Ohta's exponential distribution is the preferred distribution to be used for the selection coefficients. His (Gillespie (1991) [19]) argument goes as follows: Suppose at a particular locus, the absolute fitnesses of m alleles are given by the random variables $X_1 > X_2 > \cdots > X_m$, which are ordered m random variables drawn independently from some probability distribution. It is possible to find a sequence of numbers a_n and b_n such that the distribution $Z_n = \frac{X_1 - a_n}{b_n}$ converges to the extreme value distribution $\lim_{n\to\infty} P(Z_n < z) = exp(-e^{-z})$. The fact that the limiting distribution does not depend on the X_i is reminiscent of the Central Limit Theorem. Thus, as $m \to \infty$, the extreme value theory tells us that the distribution of s approaches an exponential distribution rather than a gamma, no matter what the distribution of the X_i. Any fast decreasing density for the selective advantages of beneficial mutational effects, would also, like the exponential, assume that there are many more beneficial mutations of small effect than of large effect.

Fisher (1930) [16] outlined a view of evolutionary adaptation in terms of intuitive, geometrical considerations. Fisher illustrated how adaptation is determined by a number of different features of an organism. An organism was described as having n quantitative traits. These quantitative characters of an organism are then viewed as

the Cartesian coordinates in an n-dimensional "space of characters", and a particular organism, with its particular set of n characters, was then geometrically represented as a point in this space. The level of adaptation of an organism was determined from its distance from a fixed point in the $n-$dimensional character space. The closer an organism is to this fixed point, the higher is its fitness. This fixed point was thus implicitly taken as a fitness optimum. A mutation is adaptive if an individual carrying a newly arisen mutation is closer to the location of the fitness evolutionary adaptation.

Models of the process of adaptive evolution have been developed, for example, by Haldane (1927) [21], Fisher (1930) [16], Kimura (1962) [30], and Ohta (1977) [43]. Adaptive evolution is driven by the emergence of beneficial mutations and their subsequent spread due to natural selection. These models demonstrated the importance of stochastic sampling events in the establishment of beneficial mutations in evolving populations. A key difficulty in the application of these models is that many of the relevant population genetics parameters are poorly known, limiting their ability to predict general features of evolutionary dynamics in real populations. In 1991, Lenski et al. [35] measured the degree to which adaptation of independent evolving populations is associated to evolving populations as a whole. The models of adaptation assume that mutations are rare, and the fate of each beneficial mutation is decided on its own merits. Further in 1998, Gerrish and Lenski [18] modeled the fate of beneficial mutations by considering clonal interference, whereby, two or more beneficial mutations arise independently and interfere in their respective growth. Some of the main conclusions of their work include

(i) The probability of fixation of a given beneficial mutation decreases with both population size and mutation rate.

(ii) As population size or mutation rate increase, adaptive substitutions result in larger fitness increases.

(iii) The rate of adaptation is an increasing, but decelerating, function of both population size and mutation rate.

(iv) Beneficial mutations that become transiently common but do not achieve fixation because of interfering beneficial mutations are relatively abundant.

In sexual populations, beneficial mutations that occur in different lineages may be recombined into a single lineage (Peters and Otto (2003) [45]). However, in asexual populations, the clones that carry such alternative beneficial mutations compete with one another, and interfere with the expected progression of a given mutation to fixation. The idea that beneficial mutations must compete in asexual populations was originally proposed by Muller in 1932 [42]. Clonal interference is thus the phenomenon whereby the fate of the beneficial mutation is altered by the appearance of a superior alternative mutation (Atwood et al. (1951) [2], Helling et al. (1987) [24], Visser et al. (1999) [1]). Such competition between beneficial mutations slows the spread of and may even eliminate the first mutation. Asexual populations adapt to their environment by the occurrence and subsequent rise in frequency of the beneficial mutations. Clonal interference ensures that those beneficial mutations that do achieve fixation are of large effect. In 1999, Miralles et al., [41] measured the effects of clonal interference in asexual RNA virus vesicular stomatitis virus. In

large asexual population, beneficial mutations compete with each other for fixation. Recent work (Wilke (2004) [53]) shows that as the population size increases, the rate of substitution approaches a constant which is equal to the mean effect of new beneficial mutations. Wilke also shows that mean effect of new beneficial mutations is smaller than the mean effect of new deleterious mutations, and that the new beneficial mutations are exponentially distributed. The mean effect of fixed mutations grows logarithmically with the population size. Wilke derives a formula whether at a given population size, the beneficial mutations are expected to compete with each other or go to fixation.

In large asexual populations, recent work (Joseph and Hall (2004) [26], Desai et. al. (2007) [12], Desai and Fisher (2007) [11], Fogle et al. (2008) [17]) has shown that beneficial mutations can be very common. When beneficial mutations are common, many will occur before any of the mutation can fix, so there will be many different mutant lineages in the population concurrently. In asexual populations, these different mutant lineages interfere and not all can fix simultaneously (Perfeito et al. (2007) [44]; Gresham et al. (2008) [20], Kao and Sherlock (2008) [27]). Work of (Visser and Rozen (2006) [10] and Desai and Fisher (2007) [11]) for instance, analyzes dynamics of such multiple mutations and the interplay between multiple mutations and interference between clones.

1.2 Recent Developments

In the experimental population evolution studied in this thesis, the populations of bacteria *Escherichia. coli* evolve over generations with daily {growth + dilution} cycles. These daily dilutions create "bottlenecks" in the population in which a rare beneficial mutation might also be lost. An important feature of these population bottlenecks in experimental systems is their extreme regularity. Unlike stochastic environmental factors, experimental bottlenecks occur at fixed intervals. At the end of each interval, the population is reduced by a fixed dilution ratio. Levin et al. (2000) [37] studied the cumulative effects of periodic bottlenecks in such growth dilution models. The work of Gerrish and Wahl (2001) [52] establishes the probability that the beneficial mutations ultimately become extinct in a population with periodic dilutions. Periodic dilutions affect evolution, and thus increase the probability that beneficial mutations will be lost. These bottlenecks occur at random and periodic intervals, with a fixed dilution ratio. The authors, Gerrish and Wahl (2001) [52] use a discrete approach based on branching process and a continuous diffusion process, solving the Kolmogorov backward equation; in order to derive this probability that a beneficial mutation is lost in a population with periodic dilutions. These authors conclude that both approaches lead to the same extinction probability. This probability drops steeply with increasing time and decreasing selective advantage s, i.e, mutations that occur late or have low selective advantage s are unlikely to survive. Usually population size is largest just before a dilution, and more mutations occur when population size is large. Further in 2002, Heffernan and Wahl [22] explore the effects of introducing genetic drift into models of evolution of population

with periodic bottlenecks. These experiments, similar to the ones considered for this thesis, are characterized by exponential growth or logistic growth, for the bacterial populations, with periodic bottlenecks. Heffernan and Wahl [22] also determine the probability that rare mutations are eliminated from the populations due to periodic bottlenecks.

In the stochastic setup of the experimental model considered for this thesis, the number of mutations is assumed to have a Poisson distribution dependent on the size of the population. Fluctuations in bacterial populations that occur between generations cannot be modeled using a deterministic growth. Stochastic distributions, such as the Poisson distribution, for mutants in each generation, allow for random fluctuations in population sizes. These fluctuations may also sometimes eliminate a rare beneficial mutation, a process known as genetic drift. The probability that a beneficial mutation will reach fixation, in a constant size population, was first addressed by Haldane (1927) [21] and Fisher (1930) [16], using a discrete treatment based on branching process. A more general continuous solution was developed by Kimura (1957) [29]; (1962) [30]) based on Kolmogorov backward equation. Gerrish and Wahl (2001) [52], derive approximation for extinction probability that a rare mutation will eliminate from the population which undergoes periodic dilutions. But in all these cases, a weak selection is assumed with a deterministic exponential model to approximate bacterial growth. Thus the probability that a rare mutation is eliminated by population bottlenecks alone is determined ignoring other factors which may influence survival, for instance, fluctuations in bacterial population. Heffernan and Whal (2002) [22] derive fixation probabilities for mutations, with large selective advantage

as well, and thus not assuming weak selection alone. In 1967, Ewen [13] derives this probability as well, but assumes that the mutant has a selective advantage in every generation, i.e, the mutants have the same selective advantage in surviving the bottleneck as it has during the growth period before bottleneck. However, the more likely scenario in the evolution of bacterial population, is that an advantageous mutant may have a selective advantage during growth, but the periodic bottleneck will select individuals at random from a population.

Experimental and theoretical analyses of evolving viral populations have been able to test the key predictions of adaptation models, including the distribution of beneficial mutation effects (Rokyta et al. (2005) [49]; (2008) [48]). This work has taken advantage of the small genome size, high mutation rate and large mutation effect size of viruses to isolate genotypes that were shown to have single adaptive mutations and that experienced minimal selection through competition between co-occurring lineages, to determine mutational parameters directly. In bacterial systems, however, direct estimates of mutational parameters is currently unfeasible (Rozen et al. (2002) [51]). Notably, it is technically difficult to ensure that observed fitness changes are due to single mutations, and to evaluate the effect that interference between competing mutations will have on biasing the fitness effect of the mutations that fix (Rozen et al. (2002) [51]). Beneficial mutations are very rare events and are thus difficult to observe. Bacterial populations seem ideal for studying beneficial mutations because these bacterial populations have large sizes and short generation times. Bacterial populations propagated in the laboratory over a relatively short period of time can undergo billions of replications. In such experiments, beneficial mutations

are sure to arise, however, become detectable when they have achieved observable frequencies in the population. This can be seen, as in, Rozen et al. (2002) [51]. To achieve observable frequency in a population, a beneficial mutation must survive both drift and clonal interference, thus creating a bias in the beneficial mutations that are observed. Some experimental designs reduce these issues, but they typically consider adaptation caused by a subset of all available beneficial mutations (Kassen and Bataillon (2006) [28]; MacLean and Buckling (2009) [39]; McDonald et al. (2011) [40]). Kassen and Bataillon (2006) [28], for instance, restricted attention to mutations with a specific phenotype, and their pleiotropic effects in different environments rather than the full spectrum of mutational effects in any specific environment. Thus considering mutations that represent a fraction of all possible mutations in a given environment. Alternatively, these parameters have generally been estimated indirectly by inferences from their effect on the linked observable markers, as seen in Imhof and Schlotterer (2001) [25], Rozen et al. (2002) [51]; Hegreness et al. (2006) [23]; Perfeito et al. (2007) [44]; Barrick et al. (2010) [5]. Details of these experiments differ, but most have in common the application of some kind of model to infer underlying evolutionary parameters from changes in the frequency of a neutral marker linked to the new arising beneficial mutation.

1.3 Outline of the Thesis

Analysis of marker divergence experiments allow estimation of an effective beneficial mutation rate, which is not possible through any direct measure of evolved genotypes.

Here, we extend the existing published results by focusing on stochastic modeling and estimation techniques for the mutational parameters in individual evolution experiments tracking the dynamics of neutral markers in the evolving populations. We focus our analysis on evolution trajectories for which emerging mutants reaching non-negligible frequencies actually have nearly identical selective advantages.

We first develop the theoretical background and study the detailed Poisson process model for the first step of the evolutionary dynamics of an asexual bacterial population evolving under the simplifying assumption that a single type of irreversible mutation is available to the population. We introduce new estimators $\hat{\mu}$ of mutation rate μ and $\hat{s}$ of mutations' selective advantage s and we precisely study the accuracy of estimators $\hat{\mu}$ of mutation rate μ and $\hat{s}$ of selective advantage s.

We apply our mutation parameters estimation techniques to six sets of experimental data collected from populations of bacteria *Escherichia coli* by T. Cooper's laboratory at University of Houston Biology Department.

We analyze the influence of key parameters of the experimental design such as the number $\mathcal{N}$ of population replicates on the accuracy of our estimates $\hat{\mu}$ and $\hat{s}$. By studying the results of six T. Cooper experiments and of another set of similarly designed experiments performed by Hegreness et al. (2006) [23], we can define a priori practical ranges for the parameters μ and s in E. Coli populations .

Our stochastic models formalizes the T.Cooper as well as the Hegreness et al. experiments. Each one of these experiments evolves $\mathcal{N}$ E. Coli replicate populations starting with cells having, except for a neutral marker, identical genotypes. Each initial population undergoes a daily dilution at the end of each daily growth period,

before being transfered to a fresh well for the next daily dilution.

Compared to previous work on the evolutionary dynamics of the initial adaptive step in an evolving asexual population (Hegreness et al. (2006) [23]; Barrick et al. (2010) [5]), we have extended both the simulation algorithm as well as the analysis of the underlying evolutionary model.

We have developed new algorithms by determining the adequate discrete approximations of the daily continuous time growth phase by dividing it into 50 time intervals in our simulations, as opposed to 12 time intervals previously used by [23]. The key probability of the effect of mutations, mainly the probability of successful bottleneck crossing, is proved to be much closer to the one prevailing in the continuous time growth process.

We have focused on developing new estimators $\hat{s}$ and $\hat{\mu}$ for two fundamental evolutionary parameters: the selective advantage s of newly arising adaptive mutations, and the rate μ at which these mutations occur. We study the asymptotic behavior of these estimators as the number of experimental populations $\mathcal{N}$ becomes large and implement intensive simulations to study the accuracy of our estimators for various values of $\mathcal{N}$.

We then develop estimators $\hat{s}$ and $\hat{\mu}$ for the case when the population undergoes a complementary sub-sampling after each daily transfer to a new well. This complementary sub-sampling is simply used to evaluate daily the color markers frequencies. A few hundreds of cells are extracted from the new well and transferred to a "plate" dedicated to frequency counting.

We study the errors on $\hat{s}$ and $\hat{\mu}$ introduced by these approximate marker frequencies

evaluations. We show how the accuracy of our estimators $\hat{s}$ and $\hat{\mu}$ is decreased due to the effect of frequency estimation by this complementary sub-sampling.

We then extend the previous model, where all mutants are assumed to have approximately the same selective advantage , to models involving multiple mutations types. In these models, when a random mutation occurs, it has a random selective advantage s, and the probability distribution of s is determined by a fixed density function $f(s)$. We have studied different parametrized models for the density function f of selective advantages, namely :

1. Model "$E(\mu, \lambda)$": the density function $f(s)$ of selective advantages is the exponential density $f(s) = \lambda exp(-\lambda s)1_{s>0}$, where λ is a positive parameter. The mean selective advantage is $\bar{s} = \frac{1}{\lambda}$.

2. Model "$EB(\mu, \lambda, a, b)$": the density function $f(s)$ of selective advantages is the exponential density $f(s) = c(\lambda)exp(-\lambda s)1_{a<s<b}$, where λ is a positive parameter. The mean selective advantage is $\bar{s} = \frac{1}{\lambda} + \frac{ae^{-\lambda a} - be^{-\lambda b}}{e^{-\lambda a} - e^{-\lambda b}}$.

3. Model "$EMP(Hist_{first}, Hist_{win})$": the density function $f(s)$ of selective advantages is generated by smoothing Empirical Histograms of experimentally observed selective advantages.

We develop new algorithms based on intensive simulations to estimate the mean selective advantage $\bar{s}$ and the occurrence rate μ of mutations. We apply these new identification algorithms to the fitting of "multiple mutation types" models to experimental data. The T. Cooper *E. coli* evolution experiments focus on six different initial genotypes. We identify the best multiple mutation model which best fits each

one of these six experiments, each of which involves 11 replicate populations.

We also present similar bacterial evolution experiments and associated models studied by Hegreness et al. (2006) [23] (HK experiments). The experimental setup used for HK experiments is similar to the TC experiments, with quite different structural design parameter values. We present the HK estimation method for the evaluation of the mean selective advantage and mutation occurrence rate. We compare the HK method to the statistical techniques we have introduced to select a "multiple mutation types" model among the 3 categories of models presented above. We then provide the accuracy of these estimators for the HK estimation technique, and the estimation technique developed in this thesis for the multiple mutation models.

CHAPTER 2

Experimental Design: *Escherichia coli* Evolution Experiments

2.1 Biological Experimental Design

The *Escherichia coli* bacterial evolution experiments that we studied and explored were carried out at Tim Cooper's laboratory at the UH department of Biology. These experiments study the genetic evolution of populations of bacteria *Escherichia Coli*. We concentrate primarily on the experiments carried out by T.C. (hereafter will be called TC experiments), but also include a previously described experiment (Hegreness et al., [23], hereafter will be called HK experiment) as a point of reference and to demonstrate the effect of a realistic range of the experimental parameters on the

application of our model.

Both the TC experiments, as well as the HK experiments, start with a number $\mathcal{N}$ of culture wells of replicate populations, containing a population of size N_0 of *E. coli* cells. Experimental populations of *E. coli* bacteria have evolved for 20,000 generations in a uniform environment. Twelve populations of these were founded from a common ancestor (Lenski et al. (1991) [35] and Barrick and Lenski (2009) [4]). These populations have evolved under the uniform environment with glucose as the density limiting resource (Lenski et al. (1991) [35], Lenski et al. (1994) [36], Lenski et al. (2000) [7]). The populations adapted to this environment by the substitution of spontaneous beneficial mutations.

In each one of the six TC experiments we have studied , the $\mathcal{N}$ replicate populations start with a distinct initial population composed of cells having identical genotypes. One of the six initial genotypes is the common ancestor Ara-1 of the 5 other initial genotypes. In the HK experiments, all populations were started from a single, different genotype MC4100 (Hegreness et al. (2006) [23]).

In each well, the initial population was composed of a single genotype, except that half of the cells were of one marker type and the other half were of another. In these experiments, an arabinose marker was used (Lenski et al. (1991) [35]). The arabinose marker has been shown to be effectively neutral under the culture conditions used in the present series of experiments (Lenski (1988) [34]). This strain of *E. coli* is considered to be strictly asexual. $L - arabinose$ (Ara^-) is mainly used as a culture medium in most of the experiments. An Ara^+ mutant was isolated from this strain (Lenski (1988) [34]). The Ara^- and the Ara^+ colonies form red and white colonies,

respectively (Levin et al. (1977) [38] and Lenski et al. (1991) [35]) on the indicator medium. In HK experiments, yellow and cyan fluorescent protein markers were used. In both the experiments, the color markers used were neutral, having no detectable effect on the individual.

The values of the structural design parameters for the TC as well as for the HK experiments are displayed in table 2.1.

2.1.1 Daily Growth

At the beginning of each daily growth period, the initial numbers of red and white cells in each well are equal to $N_0/2$, where N_0 is the size of the population in each one of the wells at the beginning. After the nutrients of a well have been exhausted, (which occurs after approximately 8 to 12 hours) the population growth stops. The daily terminal cell population size in each well, after nutrient exhaustion, is essentially fixed and equal to N_{sat}. Thus, the growth of cells in each one of the wells increases from N_0, at the beginning, to N_{sat}, at the end of the daily growth period.

2.1.2 Daily Dilution

Every 24 hours, once the population in each well has reached size N_{sat}, a subpopulation of approximate size N_0 is sampled from each one of the wells with a fixed daily dilution factor given by $D = N_{sat}/N_0$. The extracted cells are then transferred to a new well, containing fresh growth medium. This transfer step is repeated daily for all the $\mathcal{N}$ populations. These "growth + dilution" cycles above are performed daily.

Table 2.1: Structural design parameters for the TC and HK experiments

Parameter	TC experiment	HK experiment
$\mathcal{N}$	11	72
N_0	5×10^4	2.5×10^5
N_{sat}	10^7	8.25×10^8
D	200	3300

2.1.3 Complementary Sub-sampling

Once the daily "growth + dilution" cycle above has been completed, and after transfer of the diluted population to a new well, another small random sample is extracted from the N_0 cells in the new well. These complementary samples of sizes ranging between 300 and 400 are extracted from each one of the new $\mathcal{N}$ populations, and transferred onto $\mathcal{N}$ culture plates, where the cells are allowed to grow again. This complementary sub sampling is dedicated to daily estimation of color markers frequencies. On each cell plate, after a few days, the complementary subsample of 300 to 400 cells, can be inspected by a laboratory technician who determines by visual counting the frequencies of red and white cells.

2.1.4 Data Acquisition

Once the daily "growth + dilution" cycles above have been performed, and the corresponding culture plates have been inspected for color marker frequency evaluation, these frequencies are recorded and indexed by the acquisition date t, encoded as the number of days since the start of the experiment. Actually, this marker frequency recording occurs only very few days at the beginning of each experimental population

Table 2.2: Observed red and white cells daily counts for well population Pop_1 with identical initial ancestor genotype

days	1	8	15	23	25	29	32	34	36
# Red	26	138	115	42	172	425	320	300	300
# White	24	136	75	22	48	18	1	0	0

evolution.

Table 2.2 gives an example of the observed data for one well population, recording the numbers of red and white cells, and the days at which these numbers were recorded. In the HK experiments, the daily frequencies of the two cell marker types were recorded by direct fluorimetric measurements, which generate more accurate evaluations of the daily red and white cell frequencies.

2.1.5 Inferring Dynamics of Mutants

The bacterial evolution experiments described above were designed and carried out to estimate the rate and selective advantage of the newly arising beneficial mutations. We develop new algorithms to estimate these parameters directly from the observed red and white daily frequency data.

The selective advantage is essentially the basis for evolution by natural selection. It is the characteristic of an organism that enables it to survive and reproduce better than other organisms in a given environment. Thus, only the organisms best adapted to their environment tend to survive, increasing their numbers in succeeding generations while eliminating those that are less adaptive.

During population growth, and the daily "growth + dilution" cycles, mutant genotypes with various selective advantages $s > 0$ can occur with a small probability μ at each cell division. All individuals are asexual and, therefore, when a fitter genotype (as explained above, well adapted to the environment) reaches fixation (i.e. reaches a frequency close to 1) in a population, it will drive the ancestral genotype to extinction, and thus will eliminate it from the population. This will simultaneously cause the fixation of the color marker type in which it occurred. Because adaptive mutations can arise at many sites in the genome, it is usually impossible to experimentally follow their dynamics directly. For this reason, one of the goal of these experiments is to infer the underlying genotypic dynamics from the changes in the frequencies of the two marker types.

2.2 Biological Fitness Assays

Fitness of an individual or genotype is defined as its ability to survive and reproduce, and measures its contribution to the gene pool in the next generation.

In the first TC experiment, to determine *ground truth values* for the selective advantages of the winning genotypes in 10 replicate populations staring with Ara-1, and where one color marker had reached fixation, experimental fitness estimates were obtained from four evolved clones of the “winning“ marker type isolated from each of the $\mathcal{N}$ replicate populations. The fitness of each clone was measured relative to the ancestor. A green fluorescent protein (GFP) is a protein that exhibits bright green fluorescence when exposed to blue light. The fitness of each clone was measured

using a GFP expressing derivative of the ancestor as the reference strain (Lenski et al. 1991 [35]). The use of a GFP marker enabled distinguishing competing ancestor and evolved clone sub-populations by flow cytometry, a technology that allows large population samples to be screened, increasing measurement precision over previous plate-based approaches. The ability to store these bacterial populations in a nonevolving state and to maintain a strictly clonal system of propagation enables one to estimate directly the mean fitness in a particular environment. In brief, the GFP ancestor and the evolved clones were inoculated from the frozen stocks into separate wells of 2ml 96-well plates containing lysis broth (LB). All populations were grown overnight and then diluted 10^4-fold into the same environment used for the evolution experiment-Davis Minimal medium supplemented with glucose to 25ug/ml (DM25). Populations were incubated with shaking for one day to complete one growth cycle and then transferred 1:100 to the same medium to ensure they were physiologically acclimated to this environment. The next day, each evolved clone (GFP-) was individually mixed with the ancestor (GFP+) and a 1:100 dilution made into fresh DM25. Competitions were carried out over two growth cycles. At the beginning and end of each competition, samples were screened by flow cytometry to determine the fraction of GFP+:GFP- cells. To do this, a 1:10 dilution of cells was made in purified water and SYTO17, a dye that fluoresces red when bound to DNA, was added to a final concentration of 200 nM. This cycle allowed to reduce background noise by only recording events that were above a threshold red fluorescence characteristic of bacterial cells. At least 10,000 events were recorded for each sample time point. Fitness of the evolved clone was calculated relative to the ancestor as the ratio of

each strain's Malthusian parameter, estimated as $\log(F_f \times 10^4/F_i)$, where F_f and F_i are the final and initial frequencies of one cell type, respectively, and the 10^4 factor accounts for growth during the competition.

CHAPTER 3

Stochastic Model for *E. coli* Evolutionary Dynamics

3.1 Stochastic Population Growth Model

To estimate rigorously, the evolutionary parameters underlying the dynamics of new adaptive mutations in experiments of the type described above (chapter 2), we present a theoretical model which can

1. Account for the stochastic occurrence and subsequent dynamics of adaptive mutations, and

2. Relate these dynamics to the observable dynamics of the deliberately introduced colored markers that initially separate each evolving population into two distinguishable subtypes: "red" and "white" cells.

We develop an analytical model describing the dynamics of evolutionary populations in each of the $\mathcal{N}$ wells, and estimating the rate and selective advantage of the newly occurring beneficial mutation in the population.

Random Cell Splitting as a Continuous Poisson Process: The growth phase of cells in one day is due to cell splitting at random times, and we model the random process of cell splitting by a continuous Poisson process. Our model describes the dynamics of a population of a unicellular, asexual organism, for instance, *E.Coli* in terms of a cell dividing into two cells with identical genotypes, perturbed by mutations. Each cell eventually divides into two daughter cells. The genotypes of the mother and daughter cells are identical, unless a mutation occurs in one of the daughters. We begin by considering the dynamics of cell division in the absence of mutation.

The initial size of the populations is $N_0 = 5 \times 10^4$, and these cells grow until the saturation capacity $N_{sat} = 10^7$, at the saturation time t_{sat}. Typically, this growth phase duration ranges from 8 to 12 hours, in the TC experiments. During the growth phase, the population expands by a fixed factor $D = N_{sat}/N_0 = 200$. When all cells in the initial population have the ancestral genotype ("ancestor"), and in the absence of mutations, the population expands $D-$fold during this growth phase.

We assume that individual cells wait a random time between division events, and

that this time follows an exponential distribution with probability density $f(x) = \lambda \exp(-\lambda x)$ for all $x > 0$. The rate of cell division is given by $\lambda > 0$, and is determined by the genotype of the particular cell. The mean waiting time before a cell divides is then given by $1/\lambda$. We also assume that the division times for any pair of cells are independent. We ignore the effect of cell death during the daily growth phase.

The growth phase in each well ends at the (random) saturation time t_{sat} when population size reaches and stationarizes at the fixed level N_{sat}.

Discretization of Continuous Poisson Process Growth Model To simulate the growth phase using Poisson process for cell splitting, we generate a discrete approximation of this process, by dividing the daily growth period into a large number τ of equal small time intervals $J_1, \ldots, J_\tau$, and keep track of the simulated random number of cell divisions, and hence the associated mutations, in each time interval. Denote by N_k the population size at the end of interval J_k, where $k = 1, \ldots, \tau$, so $E[N_\tau] = N_{sat}$. The Poisson process assumption implies that, given N_k, the number $S(J_{k+1})$ of cells dividing during time interval J_{k+1} has a conditional Poisson distribution with conditional mean value $S_{k+1} = cN_k$ for some constant c. Hence for any lineage of cells with genotypes identical to the initial genotype, the successive mean population sizes grow exponentially, and we have $E[N_k] = N_0 F^k$, where $F > 0$ is a multiplicative factor of progenitor cells per time interval J_k. Since $E[N_\tau] = N_{sat}$, we get that

$$F = \left(\frac{N_{sat}}{N_0}\right)^{\frac{1}{\tau}} \quad = \quad D^{\frac{1}{\tau}}$$

and we have $E[S(J_{k+1})|N_k] = (F-1)N_k$.

The modeling parameters and their values, for the TC and HK experiments are displayed in table 3.1.

Table 3.1: Modeling parameters for the TC and HK experiments

Parameter	TC	HK
τ	50	12
F	1.11	1.18
s	[0.01, 0.2]	[0.01,0.2]
μ	$[2 \times 10^{-8}, 10^{-6}]$	$[2 \times 10^{-8}, 10^{-6}]$

Deterministic Approximation of Random Growth Models: Given N_k, the conditional distribution of the number $S(J_{k+1})$ cells dividing during time interval J_{k+1}, given N_k, the size of population at the end of interval J_k, has a standard deviation $std(S(J_{k+1})|N_k) = \sqrt{(F-1)N_k}$, and hence the dispersion, also called the coefficient of variation, of $S(J_{k+1})$ is defined by the ratio of standard deviation to the mean:

$$\frac{std(S(J_{k+1})|N_k)}{E[S(J_{k+1})|N_k]} = \frac{1}{\sqrt{(F-1)N_k}}$$

has a maximum $CV_{max} = 1/\sqrt{(F-1)N_0}$ reached for $k = 0$, and decreases to a minimum value $CV_{min} = 1/\sqrt{(F-1)N_{sat}}$, reached for $k = \tau - 1$. These numbers are quite small, for instance, for the TC experiments, $CV_{\max} = 0.0134$ and $CV_{\min} = 0.001$. For HK experiments, $CV_{\max} = 0.005$, and $CV_{\min} < 0.0001$.

Thus, in simulations, $S(J_k)$ can be approximated by its conditional mean $S_k = (F-1)N_{k-1}$, and N_k can be approximated by the deterministic growth formula $N_k \approx F^k N_0$, so that $S_k \approx (F-1)F^{k-1}N_0$.

We consider first the case where we have only one type of irreversible mutation available to the population, and that mutants cannot mutate further. Thus, we assume that there is single fitness locus A, with two alleles, A and a, with A for the ancestor allele. A alleles mutate *irreversibly* to a alleles with very small probability μ per cell division, and a cannot mutate back into A. These mutations are independent events. The mutant allele confers a selective advantage $s > 0$, such that mutant individuals have a faster cell division rate of $(1 + s)\lambda$ relative to A.

In a growth phase, mutants proliferate as described above, but with a stronger multiplicative growth factor per time interval J_{k+1}, given by $M = F^{1+s}$. In the presence of mutants, in each growth period, the population reaches saturation at the end of some time interval J_n where the random integer n verifies $n \leq \tau$.

Once the growth phase is completed, N_0 cells are transferred from the current wells into new wells containing fresh medium. During the growth phase, the ancestor cells have mutated into mutants, thus the current well consists of both the ancestor cells as well as the mutants. These daily dilutions introduce "bottlenecks" that reduce the probability that emerging mutants will persist in population. This daily dilution thus may eliminate a beneficial mutant present in the well.

Initially, a population contains N_0 cells, with each marker carrying half of the cells. Let $p(t)$ be the frequency of one marker. The *marker frequency ratio* $p(t)/(1-p(t))$ remains initially at 1 because marker has no effect on fitness. Most of the time, none of the occurring beneficial mutations survive the daily dilution. At some point later, after a number of successive daily {growth + dilution} cycles have completed, a beneficial mutation with a selective advantage s, will eventually cross the dilution

successfully, and thus appears in the population, typically within a single marker subpopulation. If the beneficial allele is not lost by genetic drift, it will rise in frequency, dragging along with it, the marker carried by the individual that first acquired the mutation. Thus $p(t)/(1-p(t))$ will be pushed away from 1, towards either 0 or $+\infty$. All individuals in the population are asexual, and therefore, when a fitter genotype fixes in a population, it will drive the ancestral genotype extinct and cause the fixation of the marker type on which it occurred. Thus, to follow the dynamics of the new beneficial mutation in fitter genotypes, we infer the dynamics of the beneficial mutations through their effect on the frequency of the observable markers, and hence use the dynamics of $p(t)/(1-p(t))$.We develop a theoretical framework for this in the next Section.

3.1.1 Detailed Study of Growth Phase

New mutant lineages generated: Let U_t be the total number of mutants randomly sampled at the end of day $(t-1)$ for transfer. Note that the time t is discrete and counted in days. These mutants will multiply to generate a mutant subpopulation of size $U_t M^n$ at the end of time interval J_n in day t. Then the number of divisions of ancestral genotypes in the time interval J_k, is approximately $(F-1)F^{k-1}(N_0-U_t)$. Note that the growth rate of the ancestor cells, in any time interval, is given by $F = D^{\frac{1}{\tau}}$ as the day is discretized into τ time intervals. A mutant arises with selective advantage s, and thus has an advantage of $(1+s)$ relative to the progenitor cell. The multiplicative growth factor per time interval is then given by $M = F^{1+s}$.

Let $X_k(t)$ be the number of mutants born on day t in the time interval J_k. The conditional distribution of $X_k(t)$ given the number of cell divisions $S(J_k)$ in the time interval J_k is assumed to have a Poisson distribution with mean $\mu(F-1)F^{k-1}(N_0-U_t)$. All the new mutants born on day t will multiply during the remaining of day t, thus generating new mutant lineages, and at the end of time interval J_n, the union of all these new mutant lineages form a total population of $Z_{t,n}$ mutants, where $Z_{t,n}$ is given by

$$Z_{t,n} = M^{n-1}X_1(t) + M^{n-2}X_2(t) + \cdots + MX_{n-1}(t) + X_n(t).$$

The total number of mutants $M_{t,n}$ present at the end of time interval J_n is then given by

$$M_{t,n} = M^n U_t + Z_{t,n}.$$

Given U_t, the $X_k(t)$ are independent Poisson distributed random variables. The conditional mean and variance of $X_k(t)$ given U_t are then identical, and have the form:

$$E[X_k(t)|U_t] = var(X_k(t)|U_t) = \mu(F-1)F^{k-1}(N_0 - U_t)$$

Hence, the conditional mean $z_{t,n}$ and variance $\sigma^2_{t,n}$ of $Z_{t,n}$ given U_t, is given by:

$$z_{t,n} = E[Z_{t,n}|U_t] = \mu(F-1)(N_0 - U_t)\frac{M^n - F^n}{M - F} \tag{3.1}$$

and

$$\sigma^2_{t,n} = var(Z_{t,n}|U_t) = \mu(F-1)(N_0 - U_t)\frac{M^{2n} - F^n}{M^2 - F}. \tag{3.2}$$

These expressions for $z_{t,n}$ and $\sigma_{t,n}$ imply that for all $n \leq \tau$,

$$\frac{z_{t,n}}{N_{sat}} < \mu(F-1)\frac{M^\tau - F^\tau}{D(M-F)} = \mu\left(1 - \frac{1}{F}\right)\frac{D^s - 1}{D^{s/\tau} - 1},$$

and

$$\begin{aligned}\left(\frac{\sigma_{t,n}}{N_{sat}}\right)^2 &< \frac{\mu(F-1)}{N_0}\frac{M^{2n}-F^{\tau}}{D^2(M^2-F)} \\ &\leq \frac{\mu\left(1-\frac{1}{F}\right)}{N_0}\frac{D^{2s}}{D^{(1+2s)/\tau}-1}.\end{aligned}$$

For TC experiments, we have $\tau = 50$ and the other parameters $D = 200,\ 0.01 < s < 0.2,\ \mu < 10^{-6},$ and $F = 200^{1/50} = 1.11$, and hence $\forall\ n \leq \tau$,

$$\frac{z_{t,n}}{N_{sat}} < 5 \times 10^{-6}$$

and

$$\frac{\sigma_{t,n}}{N_{sat}} < 1.25 \times 10^{-5} \tag{3.3}$$

For HK experiments, we get

$$\frac{z_{t,n}}{N_{sat}} < 7.6 \times 10^{-6}$$

and

$$\frac{\sigma_{t,n}}{N_{sat}} < 9.7 \times 10^{-6}$$

Asymptotic study of pure mutant growth If we let Q_t be the random size of the population generated at time $t < \tau$ by the initial single mutant, then we have that $Q_0 = 1$. Suppose the selective advantage of this initial mutant is given by s. To approximate by continuous time, suppose the number of discrete steps before dilution, given by τ, is large. Suppose $\tau = 500$. Then the random growth of the population Q_t has an increment $(Q_{t+1} - Q_t)$ which follows Poisson distribution with mean

$$\begin{aligned}E[Q_{t+1} - Q_t | Q_t] &= D^{(1+s)\frac{1}{\tau}}Q_t - Q_t \\ &= (D^{(1+s)\frac{1}{\tau}} - 1)Q_t\end{aligned}$$

In this random growth, there are no new mutation since we assume that mutants do not mutate again, and we consider only pure mutants in this population Q_t. When $t \to \tau$, the probability distribution of the random variable $\frac{Q_t}{E[Q_t]}$ converges to a fixed probability distribution, and all moments can then be computed explicitly. However, this distribution does not have a simple form. We can see from these calculations, that the mean and variance of this distribution are both identical to 1.

We generate simulations which start with only one mutant, and generate empirical histograms for the random variable

$$R_t = \frac{Q_\tau}{E[Q_\tau]} = \frac{Q_\tau}{D^{1+s}}.$$

Generating 1000 virtual independent one day random growths of pure mutants, we approximate this theoretical distribution by empirical distribution. Figure 3.1 displays the empirical distribution of this random variable. The empirical mean and standard deviation are quite close to 1, which confirm our theoretical findings.

Computation of the day t saturation time n_t: Next, let $n_t \leq \tau$ be the index of the first time interval J_{n_t} during which the day t population reaches the saturation size N_{sat}. Given the last inequalities (equations 3.3), we see that for each time t, given U_t, the actual value of $Z_{t,n}$ has practically no influence on the value of n_t. At the beginning of day t, the fresh medium contains U_t mutant and $N_0 - U_t$ ancestral genotypes. After k time intervals J_k, the deterministic growth of these two groups would generate $M^k U_t$ mutants and $F^k(N_0 - U_t)$ ancestral genotypes, if we ignore the

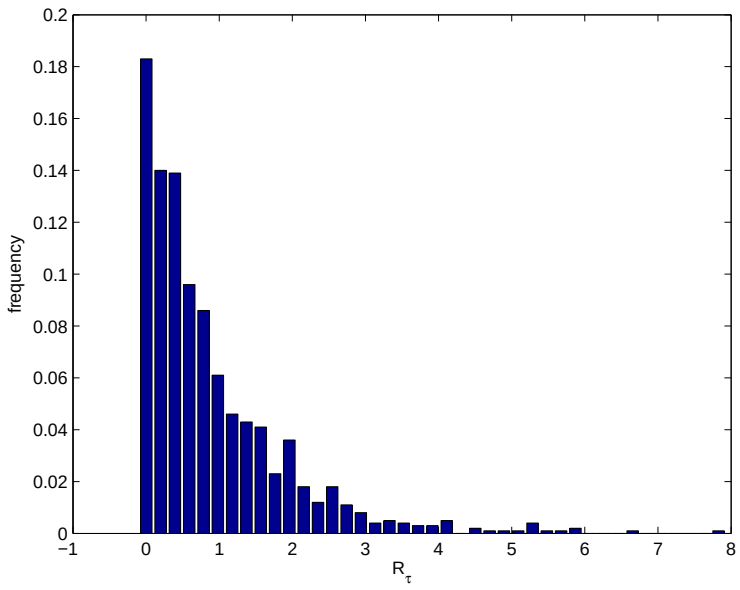


Figure 3.1: The empirical distribution of the random variable $\frac{Q_t}{E[Q_t]}$, for the simulations starting with only one initial mutant, with selective advantage 0.12, and generating pure mutant growth. The empirical mean is 0.98, and the empirical standard deviation is given by 1.02, which confirm our theoretical findings.

mutants arising at day t. The corresponding size N_k of the population verifies

$$|M^k U_t + F^k(N_0 - U_t) - N_k| < Z_{t,k}.$$

Since $Z_{t,k}/N_{sat}$ is negligible (as seen in the above calculations 3.3), we see that, given U_t, the saturation time n_t is extremely close to the unique solution k in $[1, \tau]$ of the deterministic equation

$$M^k U_t + F^k(N_0 - U_t) = N_{sat} = N_0 D. \tag{3.4}$$

This mathematical conclusion has been confirmed by simulations (see below). In the following theoretical and numerical computations, we shall therefore consider that the day t saturation time n_t is equal to the integer part of the solution of equation (3.4).

This defines an integer $n_t = g(U_t, s)$ where $g(u, s)$ is a deterministic function of $s > 0$, defined completely for all $u \in [0, N_0]$ as soon as $\tau,\ D,\ N_0$ are given. In particular, we have $g(0, s) = \tau$ for every value of s by construction.

At saturation time, the random number Z_t of mutants present in the day t cell colony, and which are descendants of new mutants born on day t is given by $Z_t = Z_{t,n_t}$. Hence, since n_t is deterministic, we obtain

$$E[Z_t|U_t] = E[Z_{t,n_t}|U_t] = z_{t,n_t} = \mu(F-1)(N_0 - U_t)\frac{M^{n_t} - F^{n_t}}{M - F}.$$

The total number of mutants $\mathcal{M}_t$ present at the saturation time in day t is given by

$$\mathcal{M}_t = M^{n_t}U_t + Z_t.$$

3.1.2 Bottleneck Crossing

Here we explain how the transition of sample from the pool of cells, from day t to day $t+1$ is a Markovian transition. At the end of day t, a random sample of size N_0 is extracted from the saturated day t cell colony, which has size

$$N_{sat} = M^{n_t}U_t + Z_t + (N_0 - U_t)F^{n_t}$$

where Z_t/N_{sat} is negligible, and n_t is such that $\frac{M^{n_t}U_t + (N_0 - U_t)F^{n_t}}{N_{sat}} \approx 1$. The number of mutants U_{t+1} present in this sample has a conditional distribution given (U_t, Z_t), which is binomial with parameters N_0 and "success" probability

$$p_s = \frac{\mathcal{M}_t}{N_{sat}} = \frac{Z_t + U_t M^{n_t}}{N_0 D}. \tag{3.5}$$

We thus see that the random vectors (U_t, Z_t) define a Markov chain. Given U_t, the conditional distributions of Z_t, $\mathcal{M}_t$, and U_{t+1} are completely determined, and we

have

$$\begin{aligned} u_t &= E[U_{t+1}|U_t] = E[p_s N_0|U_t] \\ &= \frac{1}{D}(E[Z_t|U_t] + U_t M^{n_t}) \\ &= \frac{U_t M^{n_t}}{D} + \frac{\mu(F-1)}{D(M-F)}(M^{n_t} - F^{n_t})(N_0 - U_t) \end{aligned} \tag{3.6}$$

$$\begin{aligned} var(U_{t+1}|U_t) &= E[N_0 p_s(1-p_s)|U_t] \\ &= E[U_{t+1}|U_t] - \frac{1}{N_0 D^2}(E[Z_t^2|U_t] + 2U_t M^{n_t} E[Z_t|U_t] + U_t^2 M^{2n_t}) \\ &= E[U_{t+1}|U_t] - \frac{1}{N_0 D^2}(\sigma_{t,n_t}^2 + z_{t,n_t}^2 + 2U_t M^{n_t} z_t + U_t^2 M^{2n_t}) \end{aligned} \tag{3.7}$$

where $n_t = g(U_t, s)$ is obtained by solving equation (3.4), and $\sigma_{t,n_t}, z_{t,n_t}$ are obtained as mentioned above in equations (3.1) and (3.2).

The conditional distribution of U_{t+1} given (U_t, Z_t) is a binomial distribution with mean u_t. Since $N_0 \geq 5 \times 10^4$ is large, this binomial can be well approximated by a Poisson distribution with the same mean as long as $u_t < 10$.

At the end of day t, the total mutant subpopulation has size $\mathcal{M}_t$ and is the union of two disjoint pools of mutants: the Z_t "new" mutants which gathers the lineages of all mutants produced on day t, and the $U_t M^{n_t}$ "old" mutants descended from the U_t mutants already present in the population at the beginning of day t.

The sample of N_0 individuals from the saturated day t population will contain $U_{t+1} = Y_{t+1} + K_{t+1}$ mutants, where Y_{t+1} is the member of mutants extracted from the "new" pool, and K_{t+1} is the member of mutants extracted from the "old" pool. Clearly the random variables Y_{t+1} and K_{t+1} are conditionally independent given (U_t, Z_t),

and their conditional distributions are binomial with parameters N_0 and respective "success" probabilities equal to Z_t/N_{sat} and $M^{n_t}U_t/N_{sat}$. Their conditional means are then given by

$$y_t = E[Y_{t+1}|(U_t, Z_t)] = \frac{Z_t}{D}$$
$$k_t = E[K_{t+1}|(U_t, Z_t)] = \frac{M^{n_t}U_t}{D}.$$

Since in the experiments, $N_0 \geq 5 \times 10^4$, these two binomial distributions are well approximated by Poisson distributions with means y_t and k_t, as long as, $y_t \leq 10$ and $k_t \leq 10$. An analysis of the tail of the variable Z_t shows that for TC experiments $P(Z_t/D < 2.75) > 0.99$. For HK experiments, $P(Z_t/D < 11) > 0.99$ Also, $k_t = M^{n_t}U_t/D \leq M^{\tau}U_t/D = D^sU_t$ which gives numerical bounds $k_t \leq 2.9\ U_t$ for TC experiments, and $k_t \leq 5U_t$ for HK experiments.

Hence, as long as $U_t \leq 2$, we can consider that the conditional distributions of Y_{t+1} and K_{t+1} given (U_t, Z_t) are independent Poisson distributions with means y_t and k_t. In particular, this is true as long as $U_t = 0$.

The first successful bottleneck crossing: Consider as above the evolution of a single lineage starting with N_0 progenitor cells, and submitted to daily {growth+dilution} cycles. Every day, a random sample of size N_0 is taken from a saturated population and transferred to a fresh medium. These daily dilutions introduce "bottlenecks" in the population, reducing the probability of survival of the emerging mutant in the population. The date $t = T_{bot}$ of the first bottleneck crossing is defined by $U_t = 0$ for $t < T_{bot}$ and $U_{T_{bot}} > 0$. It is the first time at which any mutants born during a growth period are transferred to fresh medium. Figure 3.2 displays examples of times T_{bot}

when the mutant first emerges in the population. This time is not directly observed from the experiments, but can be computed from the simulations. We display the histograms of T_{bot} as obtained from simulations for a fixed s and different values of μ.

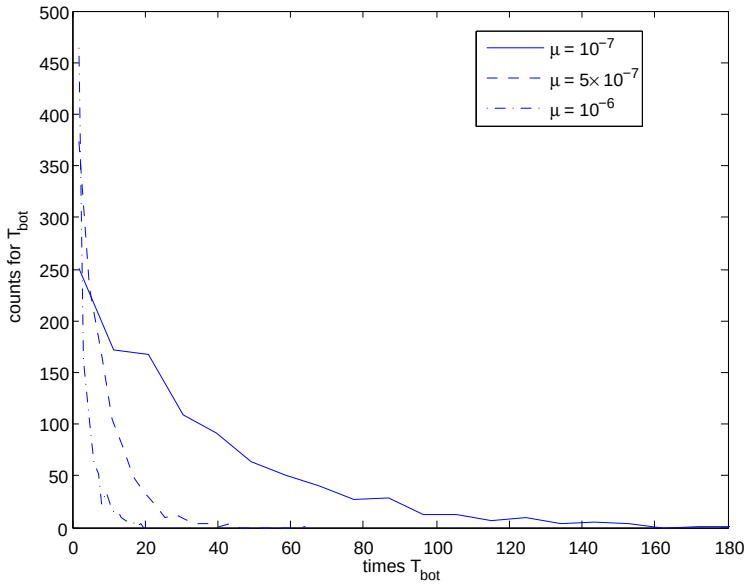


Figure 3.2: Histograms of T_{bot} for fixed $s = 0.12$ and for three different values of μ. As the rate μ increases, the emergence times for mutants become small.

As long as $t < T_{bot}$, the saturation time n_t is equal to $g(0, s) = \tau$. As just seen, as long as $t < T_{bot}$, the conditional distribution of U_{t+1} given (U_t, Z_t) is practically identical to a Poisson distribution with parameter Z_t/D. The conditional probability P_{bot} of bottleneck crossing given $U_t = 0$ does not depend on t, and is a key characteristic of the population's evolutionary dynamics. By definition,

$$1 - P_{bot} = P(T_{bot} > t + 1 | T_{bot} > t) = P(U_{t+1} = 0 | U_t = 0). \tag{3.8}$$

We will compute an explicit formula for P_{bot} below. The first bottleneck crossing

time T_{bot} has geometric distribution with parameter P_{bot}.

$$P(T_{bot} = u) = P_{bot}(1 - P_{bot})^{u-1}, \text{ for } u = 1, 2 \ldots.$$

The mean and standard deviation of T_{bot} are then given by

$$E[T_{bot}] = \frac{1}{P_{bot}}$$

and

$$std(T_{bot}) = \frac{\sqrt{P_{bot}}}{1 - P_{bot}}.$$

By applying the chain rule for conditioning, we get

$$\begin{aligned} 1 - P_{bot} &= \sum_{z \geq 0} P(U_{t+1} = 0 | Z_t = z) P(Z_t = z | U_t = 0) \\ &= \sum_{z \geq 0} e^{-z/D} P(Z_t = z | U_t = 0). \end{aligned}$$

Given $U_t = 0$, we get

$$Z_t/D = \sum_{k=1}^{\tau} \alpha_k X_k, \quad with\ \alpha_k = M^{\tau - k}/D.$$

where $X_k = X_k(t)$ have independent Poisson distributions with respective means $\lambda_k = \mu N_0 (F-1) F^{k-1}$, which implies

$$1 - P_{bot} = E[\exp(-Z_t/D) | U_t = 0] = E[\exp(-\sum_{k=1}^{\tau} \alpha_k X_k)].$$

Hence

$$1 - P_{bot} = \prod_{k=1}^{\tau} E[\exp(-\alpha_k X_k)],$$

since X_k are independent Poisson variables with respect to λ_k. The explicit Laplace transform of Poisson distributions yields then that

$$1 - P_{bot} = \prod_{k=1}^{\tau} \exp[\lambda_k \exp(-\alpha_k) - 1].$$

We can then compute $\log(1 - P_{bot})$ as

$$\log(1 - P_{bot}) = \mu N_0 (F-1) \sum_{k=1}^{\tau} F^{k-1} (e^{-M^{\tau-k}/D} - 1) = -\mu N_0 \zeta_\tau(s),$$

where

$$\zeta_\tau(s) = (F-1) \sum_{k=1}^{\tau} F^{k-1} (1 - e^{-M^{\tau-k}/D}).$$

Since $F = D^{1/\tau}$ and $M = D^{(1+s)/\tau}$, this expression becomes

$$\zeta_\tau(s) = D - 1 - (D^{1/\tau} - 1) \sum_{k=1}^{\tau} D^{(k-1)/\tau} \exp\left(-D^{s-(1+s)k/\tau}\right).$$

The true value of $\log(1 - P_{bot})$ is naturally reached only when the number τ of time intervals discretizing the daily growth period tends to ∞. Since $D^{1/\tau} - 1 \sim \log(D)/\tau$, the approximation of integrals by Riemann sums yields the explicit limit

$$\zeta(s) = \lim_{\tau \to \infty} \zeta_\tau(s) = D - 1 - \log(D) \int_0^1 D^x \exp\left(D^{s-x(1+s)}\right) dx.$$

The change of variable $\omega = D^x$ transforms this expression into

$$\zeta(s) = D - 1 - \int_1^D \exp\left(D^s/\omega^{(1+s)}\right) d\omega. \tag{3.9}$$

Note that this is a function of the selective advantage s only. The value of P_{bot} is finally given by

$$\log(1 - P_{bot}) = -\mu N_0 \zeta(s). \tag{3.10}$$

Thus

$$P_{bot} = 1 - \exp\left(-\mu N_0 \zeta(s)\right) \simeq \mu N_0 \zeta(s)$$

when $\mu N_0 \zeta(s)$ is numerically small, and hence

$$log(P_{bot}) \simeq \log(\mu) + \log N_0 + \log(\zeta(s)) \tag{3.11}$$

and hence $0.0025 \leq P_{bot} \leq 0.3162$. Figure 3.3 displays the actual values of P_{bot} for the TC experiments, as computed using the formula derived above. Also figure 4.11 displays the values of P_{bot} in comparison to the simulated P_{bot}. Figure 3.4 displays the values for the function $\zeta_\tau(s)$ as a function of s for both the TC as well as for the HK experimental parameters. Also, figure 3.5 plots the function $\zeta_\tau(s)$ as a function of s for different values of τ, for the TC experiment.

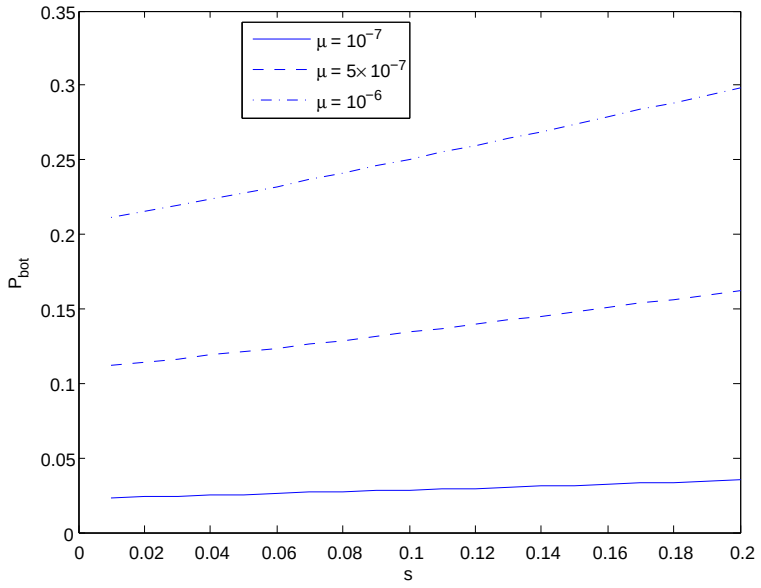


Figure 3.3: The values of P_{bot} as computed from different values of μ and s, using the function $\zeta_\tau(s)$. We see $0.0025 \leq P_{bot} \leq 0.3$. P_{bot} increases with increasing s and μ.

Comparative accuracy provided by increasing fine time discretization: The probability P_{bot} of bottleneck crossing just computed is a crucial process characteristic. When we evaluate P_{bot} by intensive process simulations after discretization of the growth phase into τ time intervals, the simulated value of $\log(P_{bot})$ will be an accurate estimator of $\log(\mu) + \log(\zeta_\tau(s)) + \log(N_0)$ instead of the desired expression

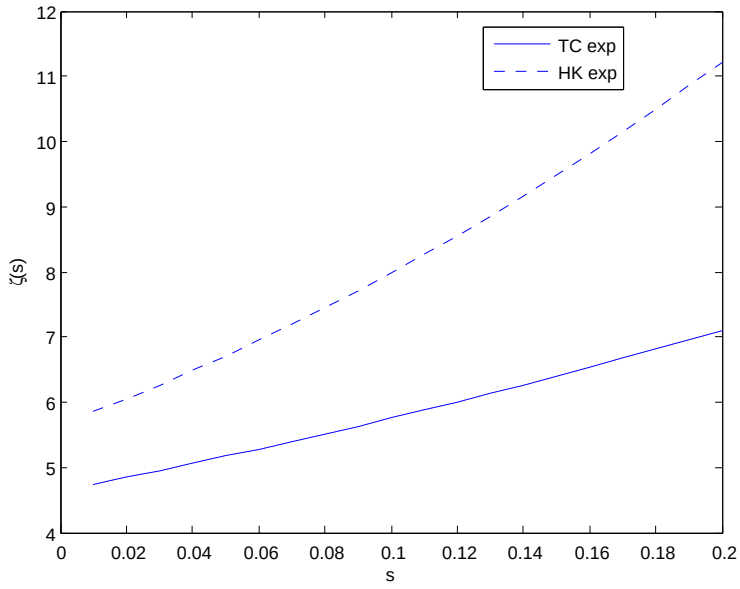


Figure 3.4: The function $\zeta_\tau(s)$, displayed for different values of s, for the TC as well as for the HK experimental parameters.

$\log(\mu) + \log(\zeta(s)) + \log(N_0)$. Hence simulation error due to time discretization is given by $|\log(\zeta_\tau(s)/\zeta(s))|$ and decreases with τ. So it is important not to choose too small a value of τ such that the accuracy of the approximation is compromised. We have calculated $\log(\zeta_\tau(s))$ numerically for a range of values of τ and s. The effect of time discretization when $\tau = 50$ is 0.052, when $\tau = 100$, this value becomes 0.026, and for $\tau = 200$, this value is 0.0129. These values are the maximum values of the accuracy statistic $|\log(\zeta_\tau(s)/\zeta(s))|$ over a range of selective coefficients $0.01 < s < 0.2$ for different values of time discretization. In view of the range of values for P_{bot} in the experiments, we expect our choice of $\tau = 50$ to yield an acceptable level of accuracy for the experiments. Figure 3.5 plots the value for $\zeta(s)$ as a function of s, for different values of τ.

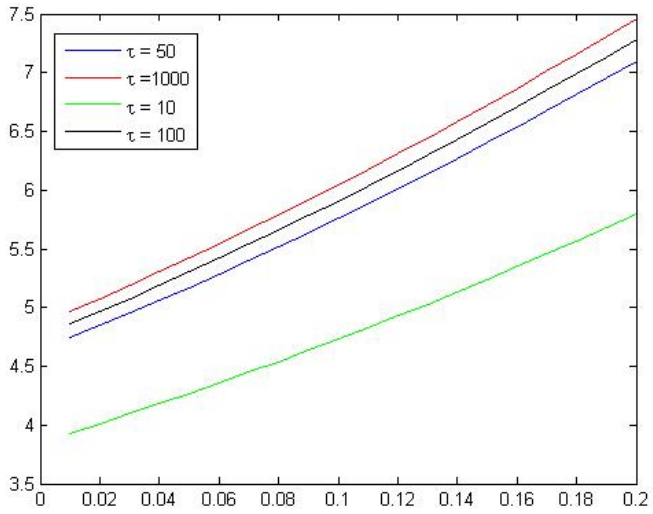


Figure 3.5: Plot for the $\zeta_\tau(s)$ as a function of s for different values of τ, for the TC experiment. The x-axis represents different s, and the y-axis plots the values $\zeta(s)$.

3.1.3 Path to Fixation

Initial increase in mutant frequency: After the first bottleneck crossing, the beneficial mutation will go to fixation with probability close to 1. With no loss of generality, we assume that $p(t)/(1-p(t)) \to \infty$, i.e, $p(t) \to 1$. However, there is also a possibility that competing lineages of mutants emerge within both the marker subpopulations before the fixation of one population, and survives the daily dilutions. Then, since all mutants have the same selective advantage in our model, the frequencies of the two markers will reach a neutrally stable equilibrium. Thus the population will contain only mutants, and marker proportions will remain stable at some frequency. Thus, there is also a possibility of stabilization of the population. Such an example is displayed later, in figure 3.9. The probability of stabilization depends on the parameter values, for the simulations. For instance, for the TC

experimental simulations, this probability of stabilization is 0.64 when $s = 0.05$ and $\mu = 2 \times 10^{-8}$, but this probability is reduced to 0.20 when $s = 0.15$ and $\mu = 3 \times 10^{-7}$. For the analysis below, we are assuming that the white cells have almost reached fixation before red mutant cells reaches a sizable frequency. We let $p(t)$ thus be the frequency of the white marker in the experiments.

Let $V_t = U_t/N_0$ be the frequency of mutants at the beginning of day t. The frequency of ancestral cells is then $1 - V_t$. Let $w_{anc}(t)$ and $r_{anc}(t)$ be the frequencies of white and red ancestral cells at the beginning of day t. Let $h(t) = \frac{w_{anc}(t)}{r_{anc}(t)}$ be the ratio of the frequencies of white and red ancestral cells at the beginning of day t. The white and red progenitor cells divide at the same rate, so that at the end of day t, there will be $w_{anc}(t)F^{n_t}$ and $r_{anc}(t)F^{n_t}$ white and red progenitor cells, neglecting the small frequency Z_t/N_{sat} of the pool of day t "new mutants", which are the descendants of lineages generated on day t by spontaneous mutants born on day t. After the bottleneck at the end of day t, the new frequencies $w_{anc}(t+1)$ and $r_{anc}(t+1)$, have conditional expectations $w_{anc}(t)F^{n_t}/D$ and $r_{anc}(t)F^{n_t}/D$, and a small conditional standard deviation (this follows from similar arguments as those developed above for Z_t/N_{sat}), at least as long as $V_t < 0.90$. Therefore, until the final approach to fixation, the ratio $h(t)$ remains quite close to 1 and $w_{anc}(t) \simeq r_{anc}(t) \simeq \frac{1-V_t}{2}$.

Since V_t denotes the frequency of mutants at the beginning of day t, hence all our V_t mutants have a white marker. Consequently, $p(t) = \frac{1-V_t}{2} + V_t = \frac{1+V_t}{2}$. The function

$$\log\left(\frac{p(t)}{1-p(t)}\right) = \log\left(\frac{1+V_t}{1-V_t}\right) \to \infty \quad as \quad t \to \infty.$$

Mutant rise towards fixation: As seen in equation (3.4), given U_t and s, the day t saturation time $n_t = g(U_t, s) = g(N_0 V_t, s)$, is the integer part of the solution k of equation

$$V_t M^k + (1 - V_t) F^k = D$$

In particular, we have approximately

$$V_t M^{n_t} + (1 - V_t) F^{n_t} \simeq D.$$

In view of equations (3.1) and (3.2), the conditional mean and variance of V_{t+1} given V_t verify

$$E[V_{t+1}|V_t] = \frac{E[U_{t+1}|U_t]}{N_0}$$

and

$$var[V_{t+1}|V_t] = \frac{var[U_{t+1}|U_t]}{N_0^2}.$$

A numerical calculation for the TC experiments, shows that the conditional dispersion or the coefficient of variation $\frac{std(V_{t+1}|V_t)}{E[V_{t+1}|V_t]}$ is smaller than 0.01 when $V_t \geq 0.176$ for all pairs (s, μ). The joint time evolution of the two variables n_t and V_t is hence driven by the following approximate, but quite accurate in practice, iterative deterministic system

$$V_t M^{n_t} + (1 - V_t) F^{n_t} \simeq D \tag{3.12}$$

$$V_{t+1} \simeq \frac{V_t M^{n_t}}{D} + \frac{\mu(F-1)}{D(M-F)}(M^{n_t} - F^{n_t})(1 - V_t). \tag{3.13}$$

Since $\frac{M^{n_t}}{D} \simeq \frac{1}{V_t} - \frac{1-V_t}{V_t}\frac{F^{n_t}}{D} > \frac{1}{V_t} - \frac{1-V_t}{V_t} = 1$, the first term in (3.13) is greater than V_t. The second term is always positive. Hence V_t increases to 1 as $t \to \infty$, and the system is valid for all $t > T$ where $T = \inf\{t|V_t \geq 0.176\}$. Notice that V_t is unobservable,

and T can only be observed through $p(t)$. Given that $p(t) = (1 + V_t)/2$, we get $T = \inf\{t : p(t) \geq 0.588\}$.

Since $\frac{M^{n_t}}{D} > 1$, the ratio R of the second term to the first term in (3.13) is bounded by

$$R < \frac{\mu(F-1)}{D(M-F)}(M^{n_t} - F^{n_t})\frac{1-V_t}{V_t}.$$

For the experiments, the range of n_t is [42,50], and the numerical evaluations show that the preceding bound for the ratio R is inferior to 1.03×10^{-4} when $V_t \geq 0.176$. Hence the second term in (3.13) is negligible when $V_t \geq 0.176$. We therefore obtain a simpler iterative deterministic system:

$$V_t M^{n_t} + (1 - V_t)F^{n_t} = D \tag{3.14}$$

$$V_{t+1} = V_t M^{n_t}/D, \quad \text{when } t > T. \tag{3.15}$$

We now study the speed of convergence of $1 - V_t$ to 0 as $t \to \infty$. Solving the system (3.14), we obtain

$$1 - V_t \simeq (1 - V_T)\frac{F^{\sum_t}}{D^{t-T}} \text{ where } \sum_t = \sum_{k=T+1}^{t} n_k.$$

For large t, we have $V_t \simeq 1$, and hence $M^{n_t} \simeq D$, which implies $n_t \simeq \tau/(1+s)$. Hence, we see that $\sum_t \simeq \frac{\tau}{1+s}(t-T)$ for large times t. A numerical study of $\sum_t$ shows that in fact $\sum_t$ is approximately a linear function of $t - T$, with slope $\tau/(1+s)$, as displayed in figure 3.6.

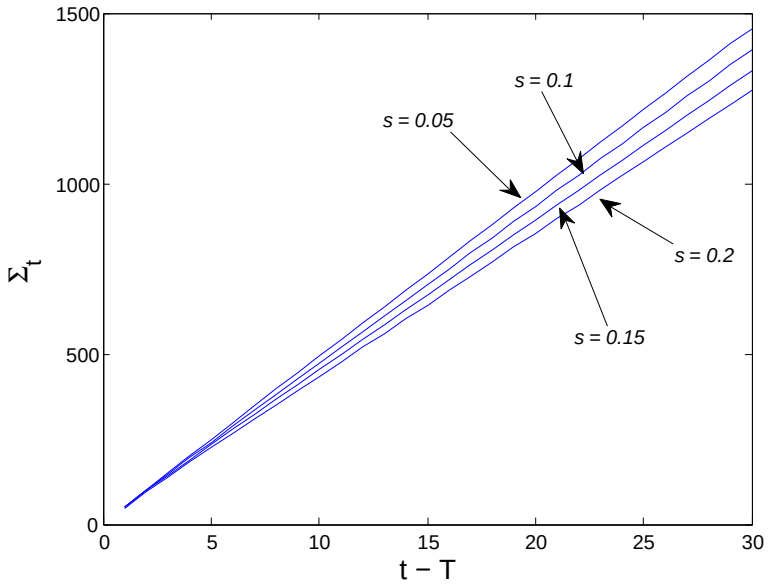


Figure 3.6: $\sum_t$ is approximately linear with respect to $t - T$.

Thus,

$$\begin{aligned} 1 - V_t &\simeq (1 - V_T)\frac{(D^{\frac{1}{\tau}})^{\frac{\tau}{1+s}(t-T)}}{D^{t-T}} \\ &= (1 - V_T)\frac{D^{\frac{t-T}{1+s}}}{D^{t-T}} \\ &= (1 - V_T)D^{\frac{t-T}{1+s}-(t-T)} \\ &= (1 - V_T)D^{-\frac{s}{1+s}(t-T)}. \end{aligned}$$

We conclude that $(1 - V_t)$ converges to 0 with exponential speed given by

$$1 - V_t \simeq (1 - V_T)/D^{\frac{s(t-T)}{1+s}}.$$

We can now study $\log\left(\frac{p(t)}{1-p(t)}\right) = \log\left(\frac{1+V_t}{1-V_t}\right)$, replacing $1 - V_t$ by the expression

just computed, to get

$$\begin{aligned} \log\left(\frac{p(t)}{1-p(t)}\right) &= \log\left(\frac{1+V_t}{1-V_t}\right) \\ &\simeq \frac{s}{1+s}(t-T)\log D + \log\frac{2}{1-V_T}. \end{aligned} \tag{3.16}$$

Taking the derivative of $\log\left(\frac{p(t)}{1-p(t)}\right)$ with respect to t, we see that the slope of $\log\left(\frac{p(t)}{1-p(t)}\right)$ is approximately linear in t for $t > T$, with slope $\frac{s}{1+s}\log D$. Note that this slope is an increasing function of s.

3.2 Examples of Evolution of Frequency

In both Hk and TC experiments, the experiments start with the same initial number of cells for both the markers (red and white), thus the frequency $p(t)$ of winner and $(1-p(t))$ of looser remain close to 0.5, until the emergence of a beneficial mutation in any one of the markers and thus increasing the frequency of that marker. This can be seen in figure 3.7. The white marker is the winner, thus $p(t)$, the frequency of white marker is close to 0.5 until the emergence of the beneficial mutation and hence causing the frequency curve to significantly deviate away from the straight line. We also present the corresponding $\log\frac{p(t)}{1-p(t)}$, curve studied above, in figure 3.8.

Figure 3.9 gives an example of the evolution of frequency of the markers when mutants appear in both the color markers with approximately same selective advantage and compete with each other, and hence causing the frequency of any one marker fluctuating around the initial 0.5.

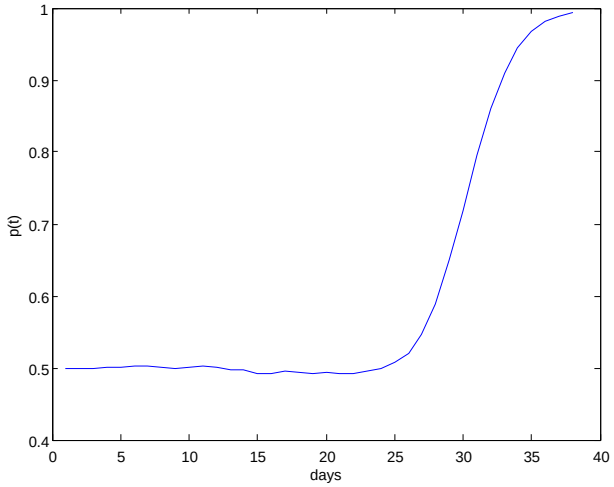


Figure 3.7: Winner = White marker. Example of evolution curve for the white frequency $p(t)$, when there is only mutation in white marker that appeared and thus the population of white marker reaches fixation. For this example, $s = 0.12$ and $\mu = 2 \times 10^{-7}$.

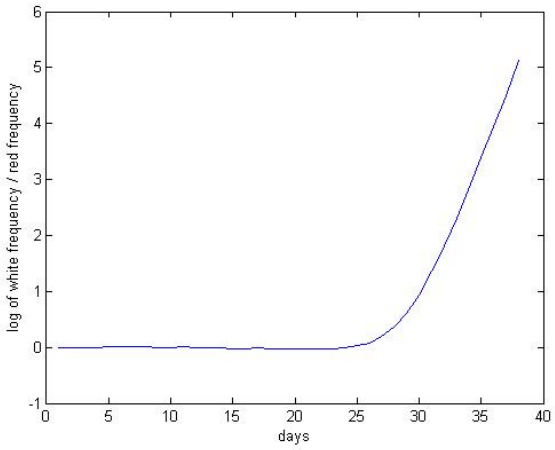


Figure 3.8: Winner = White marker. Example of evolution curve for $\log \frac{p(t)}{1-p(t)}$, when there is only mutation in white marker that appeared and thus the population of white marker reaches fixation. This is the curve corresponding to the frequency of winner plot 3.7. For this example, $s = 0.12$ and $\mu = 2 \times 10^{-7}$.

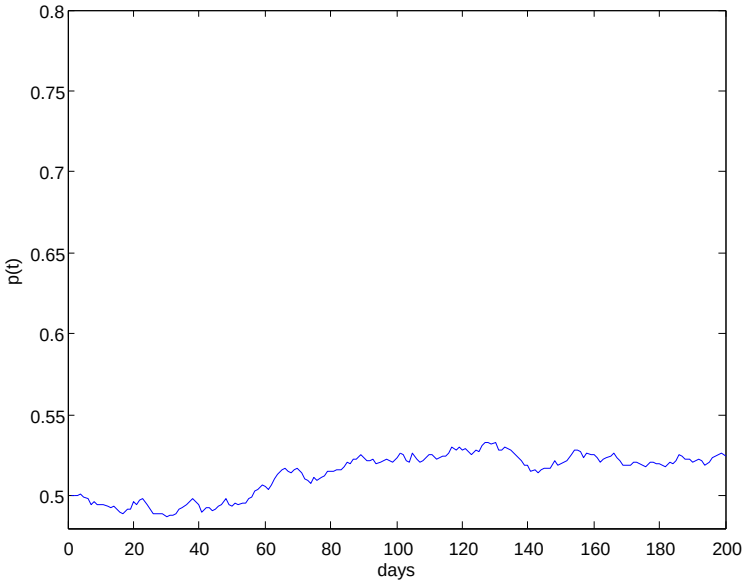


Figure 3.9: Frequency of one marker plotted against days, when mutations occur in both the marker colors and hence cause the marker frequencies to fluctuate around 0.5. For this example, $s = 0.12$ and $\mu = 3 \times 10^{-7}$.

CHAPTER 4

Parameter Estimations for Single Mutation Model

In this chapter, we introduce new methods to quantify the accuracy of the estimators. We explain the construction of confidence intervals to quantify the errors of estimation, and we then give a description of the numerical database used for these computations. After describing the methods to quantify accuracy, we will present the construction of the estimators.

As mentioned in the description of the biological experiments and the corresponding mathematical stochastic model, the parameters: N_0 (the initial population), N_{sat} (saturated population at the end of the growth day) and the dilution factor $D = N_0/N_{sat}$ are fixed. Once we fix the number τ of time intervals used to discretize

the growth periods, the multiplicative growth factor of the ancestral genotype per time interval J_k is given by $F = D^{\frac{1}{\tau}}$. The unknown model parameters to be estimated are then the selective advantage s of the mutants and the mutation rate μ of the winning mutation.

4.1 Quantifying the Accuracy of Estimators

Any estimator of s and μ must be computable as a deterministic function of the random observed experimental data, and hence has a specific probability distribution depending on the underlying unknown true parameter values of s and μ. For our stochastic model, these probability distributions have no explicit closed form expression, and we will generate good approximate empirical distributions by intensive process simulations to compute the accuracy characteristics of our estimators. Classically, when $\hat{\theta}$ is an arbitrary estimator of an unknown parameter θ, one quantifies the accuracy of $\hat{\theta}$ by several indicators, including:

- the *Bias* of $\hat{\theta}$, which is the average of $\hat{\theta} - \theta$.
- the error of estimation *Err* which is the average of $|\hat{\theta} - \theta|$.
- the average width of the 90%- confidence intervals of $\hat{\theta}$.

4.1.1 Empirical Confidence Intervals

We now present the construction of empirical confidence intervals. These confidence intervals once computed, will provide concrete view of the accuracy of the parameter estimates. For the estimators of the selective advantage, as we will see, the computation of confidence intervals is not essential since our estimators of the selective advantage have good absolute and relative accuracy. However, the estimators of the mutation rate computed below, as we will see, have much higher dispersion, and computation of the confidence intervals provides a implementable way to evaluate their accuracy. Appropriate algorithmic analysis of the simulation data will enable us to generate the empirical confidence intervals. Asymptotic confidence intervals for this estimator will be discussed in the later sections. We now present a general method for the computation of confidence intervals.

Consider an unknown model parameter $\theta \in \Theta \subset \mathcal{R}$. Let $\hat{\theta}$ be any specific estimator of the unknown model parameter θ. The probability distribution $\mathcal{F}_\theta$ of $\hat{\theta}$ also depends on θ. Fix two functions $A(u) < B(u)$ of the potential values u of $\hat{\theta}$. These functions determine the family of intervals $CI(u) = [A(u), B(u)]$. This family of intervals is called a family of "confidence intervals" at level $\alpha \in [0, 1]$, for the estimator $\hat{\theta}$, if the following is satisfied for allθ:

$$P(A(\hat{\theta}) \leq \theta \leq B(\hat{\theta})) \geq \alpha \quad \textit{when } \theta \textit{ is the true parameter.} \tag{4.1}$$

Typically, the confidence level $0 \leq \alpha \leq 1$ is fixed at a large value such as 90%. Once the two functions A and B are available and verify (4.1), the confidence interval

$$CI = CI(\hat{\theta}) = [A(\hat{\theta}), B(\hat{\theta})]$$

is computable from the available observations, via $\hat{\theta}$, and will contain the unknown true parameter value θ with probability larger than α.

Explicit closed form expressions of two functions $A(u) \leq B(u)$ defining bona fide confidence intervals are not available for the complicated processes and estimators we consider. But, by algorithmic analysis of pre-simulated data, one can actually compute accurate empirical confidence intervals for all our estimators, by the following algorithm.

Fix a desired confidence level $0.9 \leq \alpha < 1$. When θ is the true unknown parameter value, consider the two quantiles

$$\eta_- = \eta_-(\theta) \quad and \quad \eta_+ = \eta_+(\theta)$$

of $\hat{\theta}$ defined by

$$P(\hat{\theta} < \eta_-) = \frac{1-\alpha}{2} \quad and \quad P(\hat{\theta} < \eta_+) = \frac{1+\alpha}{2}.$$

For each potential value u of $\hat{\theta}$, we then define two deterministic functions $A(u) < B(u)$ by

$$A(u) = \inf\{x : \eta_-(x) < u < \eta_+(x)\}$$

$$B(u) = \sup\{x : \eta_-(x) < u < \eta_+(x)\}. \tag{4.2}$$

The confidence intervals $CI(u) = [A(u), B(u)]$ are then the bona fide confidence intervals for the estimator $\hat{\theta}$, verifying the key "confidence inequality" (4.1).

Given an arbitrary vector of observations Λ, we will then systematically compute the estimates $\hat{\theta}$, an then compute the associated confidence interval for the unknown θ as indicated above, given by

$$CI = CI(\hat{\theta}) = [A(\hat{\theta}), B(\hat{\theta})].$$

Note that CI is a deterministic function of the random estimator $\hat{\theta}$. These generic confidence intervals do verify the key "confidence inequality" 4.1.

4.1.2 Pre-computed Simulation Data Base

We now describe our pre-computed simulation data base. To pre-simulate the potential TC experiments, we fixed a rectangular grid of 1,000 pairs (s, μ), defined by 20 equally spaced values of s in $[0.01, 0.2]$, and 50 equally spaced values for μ in $[2 \times 10^{-8}, 10^{-6}]$. We then simulate 1,100 random population evolutions for each one of these 1,000 pairs of (s, μ). We have thus generated 1.1×10^6 process trajectories. Each trajectory has a duration of 200 days, and provides a sequence of 200 observations $p(t)$ for the frequency of the marker observed to be winning on the 200th day. For each fixed threshold frequency level $\beta \geq 0.55$, each simulated sequence $p(t)$ generates a value for the "post-emergence" time $T_\beta = \inf\{t : p(t) > \beta\}$. Since fixation of one marker occurs in most trajectories, we only have a few among our trajectories such that $T_\beta = \infty$. For instance, table 4.1 displays few examples of pairs s and μ such that $T_\beta = \infty$.

Table 4.1: Displaying some pairs (s, μ) for $Pty(T_\beta = \infty)$.

$s \setminus \mu$	2×10^{-7}	6×10^{-7}
0.1	0.0018	0.0136
0.15	0.0036	0.0064

For each pair (s, μ) in the grid, we thus generate an empirical sample of more than 1000 random T_β values, which accurately defines the histogram $Dis\, T_\beta(s, \mu)$ of T_β.

On the random time interval beginning at $T_{0.588}$ and ending at the fixation time $T_{fix} = T_{0.95}$, we apply linear regression to approximate the curve $\log[p(t)/(1-p(t))]$ by a regression line with slope $\hat{a}$. For each pair (s, μ) in the grid this generates a random sample of more than 1000 values of $\hat{a}$, defining the histogram $Dis\ \hat{a}(s, \mu)$ of $\hat{a}$.

The pre-computed simulation data base (SDB) stores our 1.1×10^6 trajectories of $p(t)$, as well as the histogram of both T_β and $\hat{a}$ for each one of the 1000 pairs (s, μ) in the grid.

The unknown parameter θ to be estimated is either s or μ. Once we specify further on an adequate estimator $\hat{\theta}$ of θ, we will compute its accuracy characteristics by algorithmic analysis of the pre-computed SDB.

For each pair of (s, μ) in the grid, and for each trajectory in this, we compute the value of $\hat{\theta}$. For each pair (s, μ) in the grid, this generates a random sample of more than 1000 values of $\hat{\theta}$. By averaging the corresponding 1000 values of $\hat{\theta} - \theta$ and $|\hat{\theta} - \theta|$, we obtain accurate approximations of the bias $Bias_{s,\mu}$ and the estimation error $Err_{s,\mu}$ for the estimator $\hat{\theta}$.

To calculate the empirical confidence intervals for our estimators of s and $\log \mu$, the functions A and B must first be numerically pre-computed for all potential values u of $\hat{\theta}$. For each pair (s, μ) and for each trajectory in the simulation database, we compute the value of $\hat{\theta}$, thus generating a sample of around 1000 values of $\hat{\theta}$ by repeated simulations, which provide empirical estimates of the two quantiles $\eta_- < \eta_+$ defined above. An inversion algorithm (as we will see below) is then applied to the two functions η_- and η_+ to compute the two functions A and B as determined by

(4.2). For each pair (s, μ) in the grid, we then use our sample of 1000 random values of the estimator $\hat{\theta}$, to compute 1000 confidence intervals $[A(\hat{\theta}), B(\hat{\theta})]$. We fix the confidence level at $\alpha = 0.90$. These confidence intervals will then contain the unknown value of θ for 90% of all experiments, i.e, at least 900 of these confidence intervals contain the true value θ. The average width $B(\hat{\theta}) - A(\hat{\theta})$ of these confidence intervals is a deterministic function $l(CI)$ of θ, and gives a robust evaluation for the accuracy of the estimator $\hat{\theta}$.

The three evaluations of accuracy mentioned above, tends to be small when s or μ are close to their boundaries on the grid. This is due to the boundary effect. Hence, for the following sections, we consider these three aspects on a interior of the interval of s and μ.

4.2 Parameter Estimation

In this section we develop the estimators $\hat{s}$ and $\hat{\nu}$ of s and $\log \mu$. For this section we restrict our study to considering the assumption that only one irreversible mutation is available to the population, for which we have developed the theoretical framework previously as mentioned in chapter 3. We first develop and construct the estimator $\hat{s}$ of the selective advantage s, independently of the other unknown parameter μ. We then develop the estimator $\hat{\nu}$ of $\log \mu$ using the fact that estimator $\hat{s}$ is known.

4.2.1 Selective Advantage

We first develop a preliminary estimate $\hat{s}_{pr}$ of s and then transform it to form the final unbiased estimator $\hat{s}$ of s. As mentioned above, we apply linear regression on a random time interval beginning $T_{0.588}$ and ending at $T_{fix} = T_{0.95}$, to approximate the curve $\log[p(t)/(1-p(t))]$ by a regression line with slope $\hat{a}$. By equation (3.16), we see that the curve $\log \frac{p(t)}{1-p(t)}$ is approximately linear in t for $t > T$, with slope $\frac{s}{1+s} \log D$. Thus we have

$$\begin{aligned} \hat{a} &= \frac{s}{1+s} \log D \\ \hat{a}(1+s) &= s \log D \\ s(\log D - \hat{a}) &= \hat{a} \end{aligned}$$

And thus we define a preliminary estimate $\hat{s}_{pr}$ of s by,

$$\hat{s}_{pr} = \frac{\hat{a}}{\log D - \hat{a}}. \tag{4.3}$$

Figure 4.1 displays an example of the curve of $\log \frac{p(t)}{1-p(t)}$. Displayed in red is the linear part, and the black dotted line is the regression line obtained after applying linear regression on this part.

For each of our parametric grid of 1000 pairs (s, μ), the empirical distribution $Dis\ \hat{a}(s, \mu)$ allows us to evaluate accurately the empirical mean and median of this estimator $\hat{s}_{pr}$. The median of $\hat{s}_{pr}$ is a function of s and μ, denoted by $G(s, \mu)$. But $G(s, \mu)$ turns out to be practically independent of μ, as displayed in figure 4.2.

This function is close to a linear function of s for fixed μ. Since $G(s, \mu)$ is practically independent of μ, a linear regression of $G(s, \mu)$ with respect to s generates

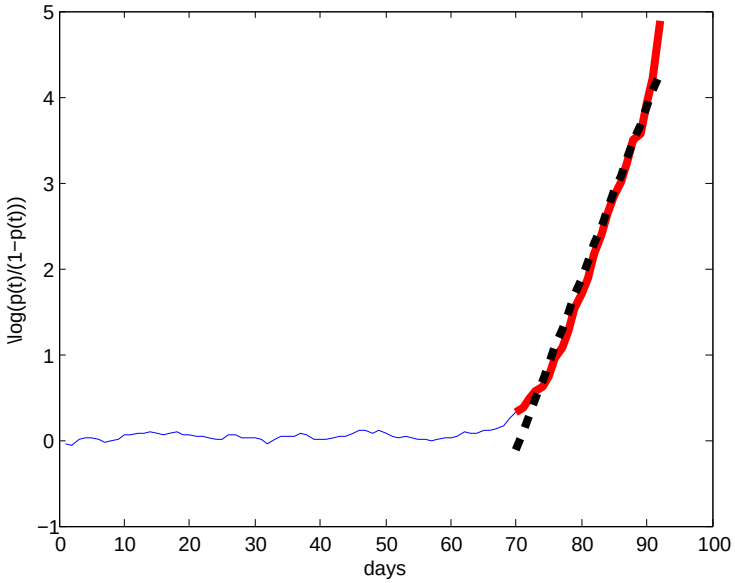


Figure 4.1: An example of the curve $\log \frac{p(t)}{1-p(t)}$, which becomes linear after the first strong deviation from the straight line in the curve of frequency of winner $p(t)$. Applying linear regression after this time and displaying the regression line in dotted black line.

the linear approximation

$$G(s) = G(s, \mu) \simeq 0.787s - 0.003.$$

The estimate $\hat{s}_{pr}$ is biased because the derivative $G(s, \mu)$ with respect to s is close to $0.787 < 1$. We hence generate a new unbiased estimator $\hat{s}$ of s by

$$\begin{aligned} \hat{s} &= G^{-1}(\hat{s}_{pr}) = (\hat{s} + 0.003)/0.787 && (4.4) \\ &= \left(\frac{\hat{a}}{\log D - \hat{a}} + 0.003 \right) /0.787 && (4.5) \\ &= \frac{1.26\ \hat{a} + 0.002}{5.29 - \hat{a}}. && (4.6) \end{aligned}$$

Since $G(s) = G(s, \mu)$ is an increasing function of s, the median of the estimator $\hat{s}$ is $G^{-1}(\cdot)$ evaluated at the median $\hat{s}_{pr}$, $G(s)$. So, the median of $\hat{s}$ is $G^{-1}[G(s)] = s$ and

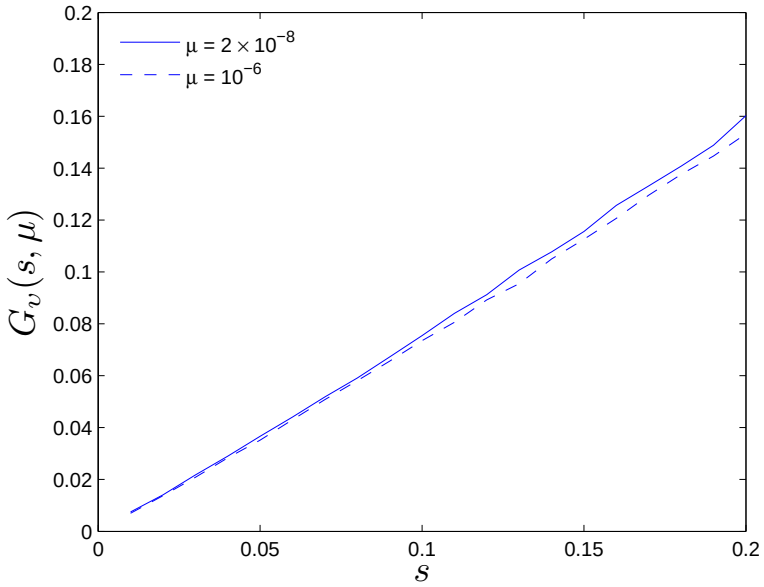


Figure 4.2: The empirical median of $\hat{s}_{pr}$ as a function of s for two extreme values of μ, is independent of μ and not equal to s.

$\hat{s}$ is an unbiased estimator of s.

Let $Dis\ \hat{s}(s,\mu)$ be the empirical histogram of $\hat{s}$ for fixed s and μ generated by simulations, using the SDB. Figures 4.3–4.4 show that $Dis\ \hat{s}(s,\mu)$ is centered at s and is practically independent of μ.

The median of $\hat{s}$ is a function of s and μ, denoted by $MED(s,\mu)$. We plot $MED(s,\mu)$ as a function of s for two extreme values of μ, for $\mu = 2\times10^{-8}$ and 10^{-6} in figure 4.5. The slopes of these two lines are both close to 1, confirming that $\hat{s}$ is an unbiased estimator.

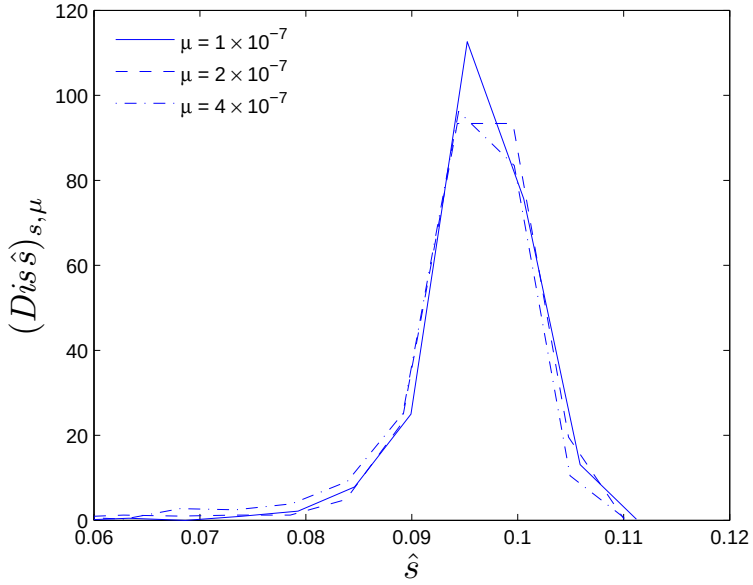


Figure 4.3: Empirical histograms of $\hat{s}$ for 3 values of μ are almost identical for a fixed $s = 0.1$.

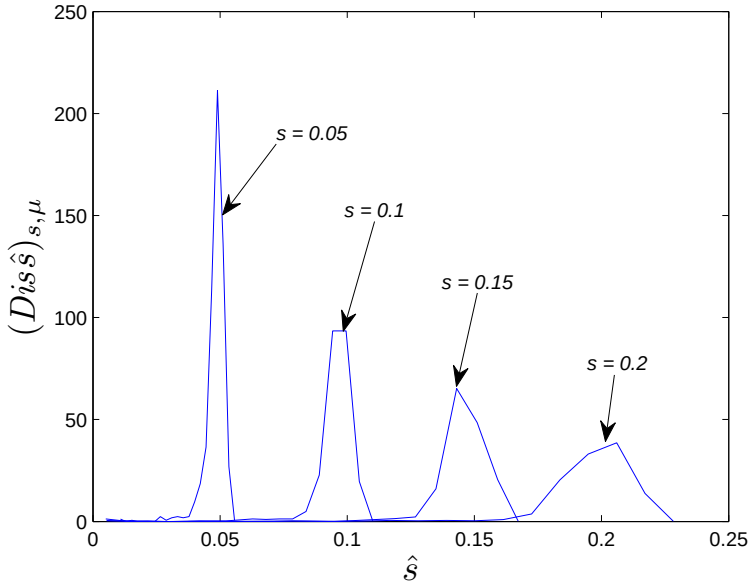


Figure 4.4: Empirical histograms of $\hat{s}$ are centered at s as displayed for 4 values of s and for a fixed $\mu = 2 \times 10^{-7}$.

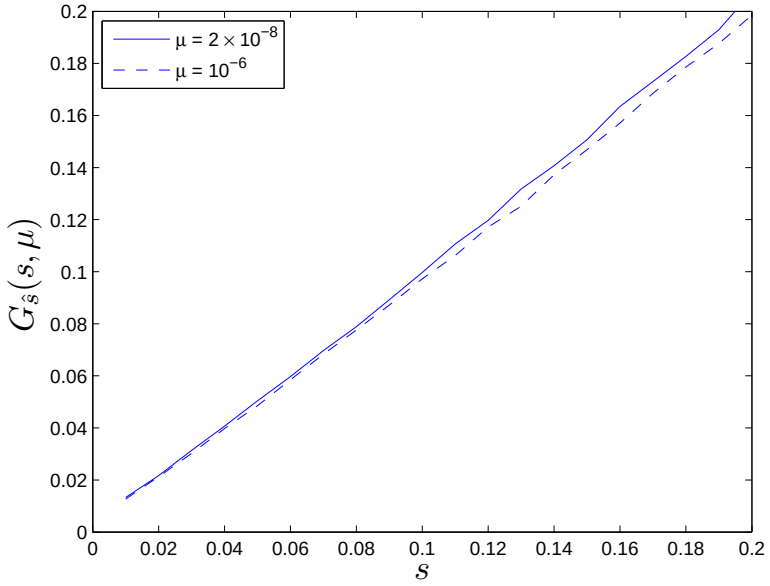


Figure 4.5: Empirical median of $\hat{s}$ is approximately s, and hence confirming the estimator $\hat{s}$ is unbiased.

4.2.2 Accuracy of the Selective Advantage Estimator

We characterize the accuracy of the estimator $\hat{s}$ by the indicators described above in Section 4.1, the average bias $Bias_{s,\mu}$, the average estimation error $Err_{s,\mu}$, the average width $l(CI)_{s,\mu}$ of the confidence intervals at confidence level 90%. In our range of parameters, as explored by the finite grid of pairs (s, μ), we have numerically verified that the histograms of the estimator $\hat{s}$ of the selective advantage of the winner, as shown above, are practically independent of the unknown μ value. Thus our accuracy characteristics $Bias$, Err, and $l(CI)$ depend essentially on the true value of s. Figures 4.7, 4.8 and 4.9 display the plots for the three accuracy characteristics: Err, $Bias$ and the mean length of CI, of estimation, for a single population ($\mathcal{N} = 1$). Also, figure 4.6 displays an example of the quantile curves and the computation of CI for s using the inversion algorithm as detailed above in Section 4.1.1.

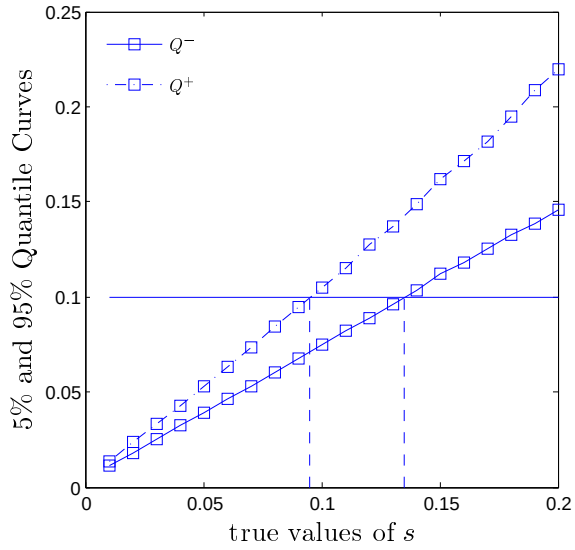


Figure 4.6: Example of Quantile curves and confidence intervals: The solid horizontal line gives a value of $\hat{s} = 0.1$, the absicca of the intersection of $\hat{s} = 0.1$ to the quantile curves give the lower and upper limit of the confidence interval.

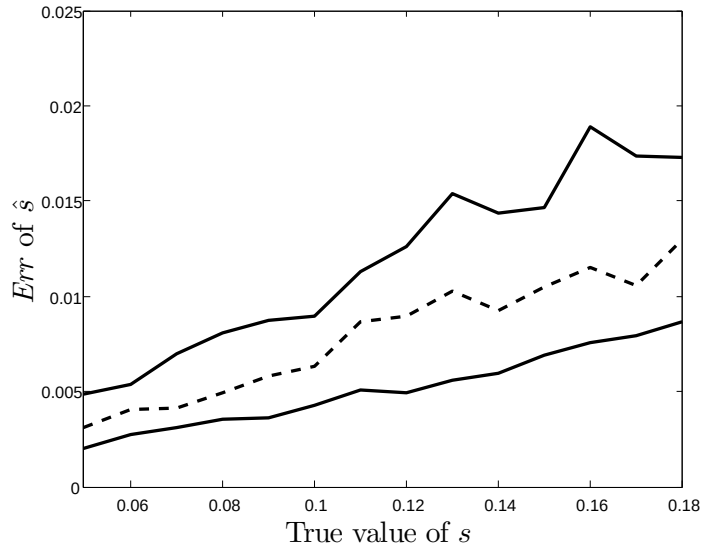


Figure 4.7: The accuracy indicator Err of estimation for $\hat{s}$ is plotted as a function of s for fixed μ, based on an observation of a single population ($\mathcal{N} = 1$). The solid lines display the Err for two extreme range values for $\mu = 2 \times 10^{-8}$ (bottom solid line) and for $\mu = 10^{-6}$ (top solid line). The dotted line displays the Err for a mid-range values of $\mu = 5 \times 10^{-7}$.

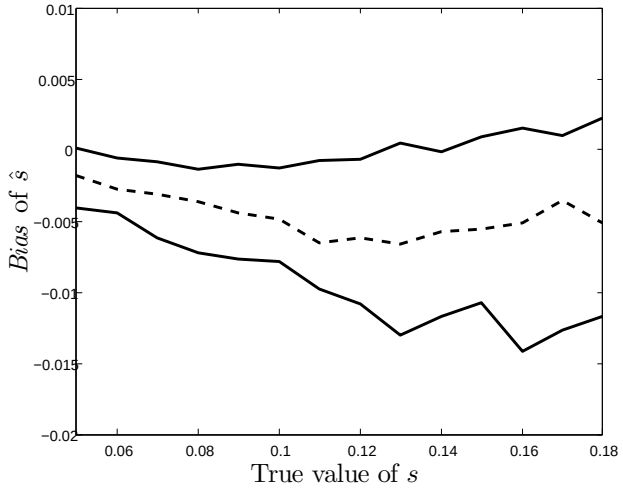


Figure 4.8: The accuracy indicator *Bias* of estimation for $\hat{s}$ is plotted as a function of s for fixed μ, based on an observation of a single population ($\mathcal{N} = 1$). The solid lines display the *Bias* for two extreme range values for $\mu = 2 \times 10^{-8}$ (bottom solid line) and for $\mu = 10^{-6}$ (top solid line). The dotted line displays the *Bias* for a mid-range values of $\mu = 5 \times 10^{-7}$.

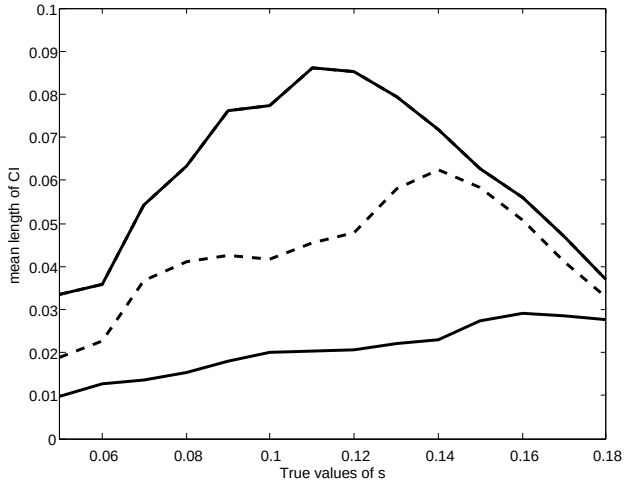


Figure 4.9: The mean length of CI of estimation for $\hat{s}$ is plotted as a function of s for fixed μ, based on an observation of a single population ($\mathcal{N} = 1$). The solid lines display the curves for two extreme range values for $\mu = 2 \times 10^{-8}$ (bottom solid line) and for $\mu = 10^{-6}$ (top solid line). The dotted line displays the mean length of CI for a mid-range values of $\mu = 5 \times 10^{-7}$.

Figures 4.7 and 4.8 show that the average bias and the average error of estimation are always quite small and always remain inferior to 0.02. The impact of μ on the bias and error of estimation is quite weak, but increases as s increases. Thus, the estimator $\hat{s}$ is an unbiased estimator regardless of the value of μ. The average width of the 90% confidence interval remains inferior to 0.09. Confidence intervals naturally have smaller width when the true s is close to the boundary of the $s-$ values range due to the boundary effect.

These numerical results, as displayed in figures 4.7, 4.8, and 4.9, show that our unbiased estimator $\hat{s}$ based on an observation of a single population is already quite accurate, but the accuracy is improved by experiments with $\mathcal{N}$ populations, providing $\mathcal{N}$ trajectories of the $\log(p(t)/(1-p(t)))$, each of which yields an independent value of $\hat{s}_j$ with $j = 1, \ldots, \mathcal{N}$. The estimate of s is then given by

$$\hat{s}(\mathcal{N}) = median\ \hat{s}_j, \qquad for\ \ j = 1, \ldots, \mathcal{N}.$$

By the Law of Large Numbers, $\hat{s}(\mathcal{N})$ converges to the $median\ \hat{s} = s$ when $\mathcal{N} \to \infty$. The estimation error of $\hat{s}(\mathcal{N})$ is approximately $Err/\sqrt{\mathcal{N}}$. For the TC experiments mentioned above, $\mathcal{N} = 11$. The error of $\hat{s}(11)$ remains inferior to $0.02/\sqrt{11} = 0.006$.

4.2.3 Logarithmic Mutation Rate

The logarithm of the mutation rate, $\nu = \log \mu$, is related to the logarithm of the conditional probability P_{bot} of bottleneck crossing on day t given that there were no mutants at the beginning of that day (3.11). In this Section we construct and study the estimator $\hat{\nu}$ of the logarithmic mutation rate. Since we considered the grid $\mu = 10^{-8}$ to $\mu = 10^{-6}$ for μ, the range of values for $\nu = \log \mu$ is $(-18.42, -13.82)$. We fix a discretization of this range by a grid GR of 13 potential values of ν:

$$GR = \{-17.73, -16.12, -15.53, -15.16, -14.89, -14.68,$$
$$-14.51, -14.36, -14.23, -14.12, -14.01, -13.92, -13.84\}.$$

The exponentials of the values in GR are evenly spaced. Given $\mathcal{N}$ trajectories for the marker frequencies, the unknown selective advantage s can be estimated by the method detailed above. Since our estimator $\hat{s}$ is quite accurate, and does not require the previous knowledge of μ, we treat the estimate $\hat{s}$ as the true value of the unknown parameter s and develop a method to then estimate ν.

The distribution of the first bottleneck crossing time T_{bot} is unbiased and $E[T_{bot}] = \frac{1}{P_{bot}}$. Since the logarithm of the mutation rate is related to the logarithm of the conditional probability P_{bot}, after estimating s, we could obtain an estimate of ν using equation (3.11) if P_{bot} is estimated. However, T_{bot} cannot be directly observed in the TC experiments, and hence we first develop an algorithm using the maximum likelihood to estimate the unobservable T_{bot} using the random observed time T_β which is the first time the frequency $p(t)$ of winning genotype exceeds a fixed threshold $\beta \geq 0.55$. For the calculations below, we select $\beta = 0.55$.

The conditional probability distribution of T_{bot} given $T_\beta = t_\beta$ verifies for all pairs $t < t_\beta$

$$\begin{aligned} P(T_{bot} = t | T_\beta = t_\beta) \cdot P(T_\beta = t_\beta) &= P(T_{bot} = t, T_\beta = t_\beta) \\ &= P(T_\beta - T_{bot} = t_\beta - t, T_{bot} = t). \end{aligned}$$

The waiting time from the time the mutant emerges and to the time it reaches fixation, $T_{wait} = T_\beta - T_{bot}$ is independent of T_{bot}, and since T_{bot} follows geometric distribution with parameter P_{bot},

$$P(T_{bot} = t) = (1 - P_{bot})^{t-1} P_{bot},$$

and hence we obtain,

$$\begin{aligned} P(T_{wait} = t_\beta - t, T_{bot} = t) &= P(T_{wait} = t_\beta - t) \cdot P(T_{bot} = t) \\ &= P(T_{wait} = t_\beta - t)(1 - P_{bot})^{t-1} P_{bot}. \end{aligned}$$

Thus, the conditional distribution of T_{bot} given $T_\beta = t_\beta$ is given by

$$P(T_{bot} = t | T_\beta = t_\beta) = \frac{P(T_{wait} = t_\beta - t)(1 - P_{bot})^{t-1} P_{bot}}{P(T_\beta = t_\beta)}. \tag{4.7}$$

Given $T_\beta = t_\beta$, the maximum likelihood estimate $\hat{T}_{bot}$ of T_{bot} is obtained by maximizing in t, the above equation (4.7), that is:

$$\hat{T}_{bot} = \arg \max_{t | 0 \le t \le t_\beta} P(T_{wait} = t_\beta - t)(1 - P_{bot})^{t-1}. \tag{4.8}$$

Since s is already estimated and considered known, then for each fixed potential value u of the unknown parameter $\nu = \log \mu$, the corresponding theoretical value of P_{bot} becomes only a function of u, since this value can be computed using (3.8) and

(4.8).

The unknown probability distribution of T_{wait}, for a potential value u of ν, can be approximated by the empirical distribution obtained from the simulations of 1100 evolution process trajectories associated with each potential value $u \in GR$.

Thus for each potential value u, and for each integer t_β, by using the maximum likelihood approach we can compute the value of the estimator $\hat{T}_{bot}$. This estimate thus depends on the potential value u of the unknown parameter ν, and on the observed value t_β of T_β. Figure 4.10 displays example of distribution of T_β for a fixed value of $s = 0.12$ and for three different values of μ.

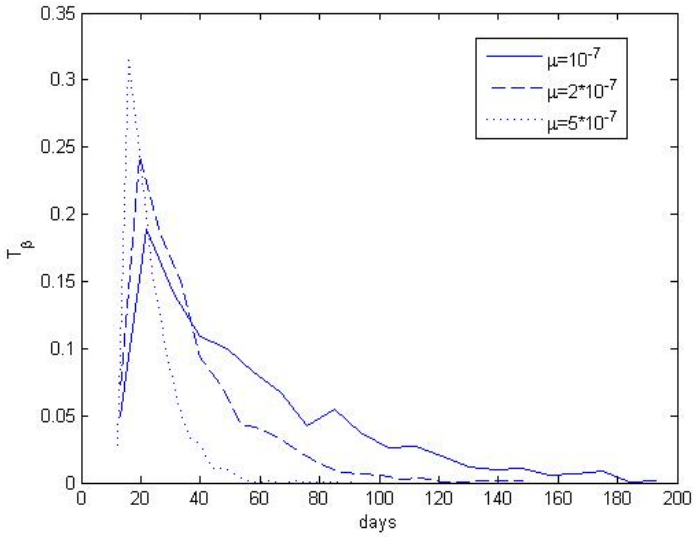


Figure 4.10: Distribution of T_β for fixed $s = 0.12$ and three different values of μ, when $\beta = 55\%$.

Let u be any fixed value in the grid GR, and assume that the unknown parameter $\nu = \log \mu$, has the true value u. The $\mathcal{N}$ observed population trajectories, generate $\mathcal{N}$

observed values of T_β, denoted by $T_\beta^1, \ldots, T_\beta^{\mathcal{N}}$. To each of these values of T_β^i, we can then apply the above maximum likelihood approach to compute the corresponding $\mathcal{N}$ estimates of $\hat{T}_{bot}^i$, for $i = 1, \ldots, \mathcal{N}$. The average of these $\mathcal{N}$ estimates gives the estimate of P_{bot}, namely,

$$\hat{P}_{bot}(\mathcal{N}) = \frac{1}{\bar{T}_{bot}(\mathcal{N})} \approx \frac{1}{E[T_{bot}]}. \tag{4.9}$$

where

$$\bar{T}_{bot}(\mathcal{N}) = \frac{1}{\mathcal{N}} \sum_{i=1}^{\mathcal{N}} \hat{T}_{bot}^i.$$

Figure 4.11 displays the probability P_{bot} as computed from the explicit formula derived above in chapter 3, and compares it to the simulated probability P_{bot}, along with the 95% confidence intervals. From simulations, the actual times T_{bot} at which the mutants emerge and survive a bottleneck successfully, can be determined, and thus by using equations 3.8, we compute the values P_{bot} empirically, using the empirical times for the emergence of mutants in the trajectories.

A crude estimate of ν, can be deduced from equation (3.11),

$$\nu = \log \mu \approx \log P_{bot} - \log N_0 - \log \zeta(s).$$

Thus $H(u)$ is given by

$$H(u) = -\log(\bar{T}_{bot}(\mathcal{N})) - \log N_0 - \log \zeta(s). \tag{4.10}$$

Note that this numerical value of $H(u)$ is a function of the potential value u we had temporarily fixed as a hypothetic value for the unknown parameter ν. The value $H(u)$ is not an estimator per se. But we expect $H(u)$ to be quite close to u when the

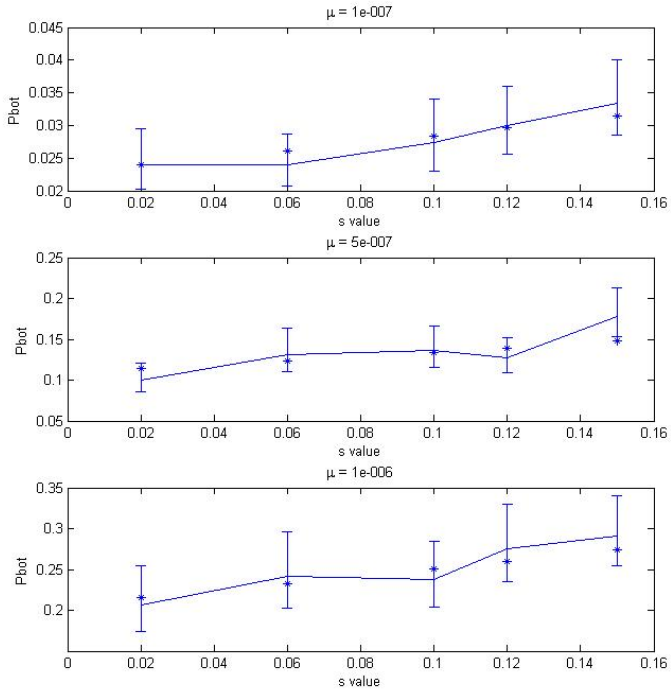


Figure 4.11: Comparison of simulated (along with 95% CI) and actual P_{bot} values, for three fixed values of μ, and for different s. For all values of μ, the actual P_{bot} is contained inside the 95% CI of the simulated $\hat{P}_{bot}$. From above, we have $\hat{P}_{bot} = \frac{1}{E[T_{bot}]}$.

potential value u is equal to the true unknown value of the parameter ν. Thus, to discover the unknown value ν, a natural approach is to determine, in the finite grid GR, which potential value u actually minimizes the absolute value $|H(u) - u|$. This minimizing u is the intermediary estimator of ν, and is denoted by $\hat{\nu}_{int}$, defined by

$$\hat{\nu}_{int} = arg \min_{y \in GR} |H(y) - y|. \tag{4.11}$$

The algorithmic computation of $\hat{\nu}_{int}$ only requires knowledge of the observed values of the $\mathcal{N}$ times $T_\beta^1, \ldots, T_\beta^{\mathcal{N}}$, and hence $\hat{\nu}_{int}$ is a bona fide estimator of ν.

For each pair (s, μ), our SDB stores an empirical histogram $DisT_\beta(s, \mu)$ for T_β (figure

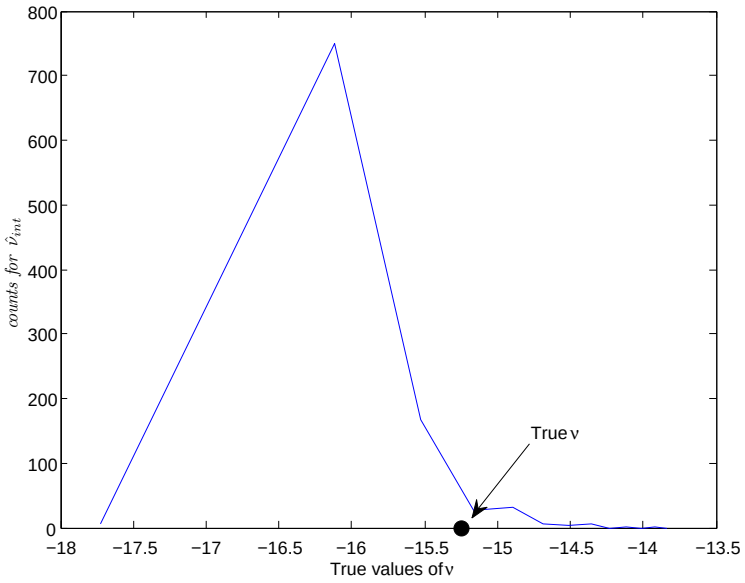


Figure 4.12: The empirical histogram of the estimator $\hat{\nu}_{int}$ when $s = 0.10$, $\nu = -15.24$ and $\mathcal{N} = 11$. We see that the estimator $\hat{\nu}_{int}$ tends to underestimate the true value ν.

4.10). Fix s and the number $\mathcal{N}$ of populations. For each $\nu = \log \mu \in GR$, we launch repeated independent random sampling from the histogram $DisT_\beta(s, \mu)$, in order to generate 1,000 lists of $\mathcal{N}$ random values of T_β, denoted $T_\beta^1, \ldots, T_\beta^{\mathcal{N}}$. Applying the estimation algorithm mentioned above to each one of these 1000 lists we generate a sample of 1,000 random values of $\hat{\nu}_{int}$, which provides an accurate empirical histogram $Dis\ \hat{\nu}_{int}(s, \nu, \mathcal{N})$ for the estimator $\hat{\nu}_{int}$. Figure 4.12 displays an example of empirical histogram $(Dis\hat{\nu}_{int}(0.10, 1.2{*}10^{-7}, 11))$ for the estimator $\hat{\nu}_{int}$ for $s = 0.10, \nu = -15.93$ and $\mathcal{N} = 11$.

4.2.4 Asymptotic Confidence Interval for ν for Large $\mathcal{N}$

Assume that the first emergence times T^i_{bot}, of mutants, for the number of populations, $\mathcal{N}$, is actually observed. Then by equation (4.10), we see that the preliminary estimator H of ν is a deterministic smooth function (call it g, where $g(x) = -\log(x) - \log\zeta(s) - \log N_0$) of the average times $\bar{T}_{bot}(\mathcal{N})$. Since $\zeta(s)$ and N_0 are fixed, the function $g(x)$ has two continuous derivatives $g'(x) = -1/x$ and $g''(x) = 1/x^2$.

When $\mathcal{N} \to \infty$, the distribution of $\bar{T}_{bot}(\mathcal{N})$ is asymptotically Gaussian by the Central Limit Theorem, and as seen in chapter 3 previously,

$$E[T_{bot}] = \frac{1}{P_{bot}}, \quad and \quad std[T_{bot}] = \frac{\sqrt{P_{bot}}}{(1-P_{bot})\sqrt{\mathcal{N}}}.$$

Since, g is a smooth deterministic function as above, it preserves asymptotic normality, (see, for instance, Azencott, 1980 [3]) and hence the estimator H must be asymptotically Gaussian as $\mathcal{N} \to \infty$, with mean $g(1/P_{bot}) = \nu$ and standard deviation $|g'(1/P_{bot})|std[T_{bot}] = \frac{P_{bot}^{3/2}}{(1-P_{bot})\sqrt{\mathcal{N}}}$. This means in particular that the estimator H converges to the true value ν as the number of populations $\mathcal{N} \to \infty$, and that the size of the estimation error is proportional to $\frac{1}{\sqrt{\mathcal{N}}}$.

Asymptotic confidence intervals for ν based on the estimator H can then be calculated theoretically using the fact that for large $\mathcal{N}$, the random variable

$$\sqrt{\mathcal{N}}(H(\nu) - \nu)(1 - P_{bot})P_{bot}^{-3/2}$$

is approximately a standard Gaussian random variable with mean 0 and variance 1. For standard Gaussian variables. the confidence intervals at 95% confidence level are

centered at 0 and have a half-width 1.96, so for large number of populations $\mathcal{N}$, we have

$$Prob\left[H(\nu) - \frac{1.96P_{bot}^{3/2}}{(1-P_{bot})\sqrt{\mathcal{N}}} < \nu < H(\nu) + \frac{1.96P_{bot}^{3/2}}{(1-P_{bot})\sqrt{\mathcal{N}}}\right] \sim 0.95. \qquad (4.12)$$

For the experiments considered, P_{bot} is often small, and these asymptotic confidence intervals have a width inferior to $2/\sqrt{\mathcal{N}}$, which is as good an asymptotic bound as one could theoretically expect, but which only becomes effectively valid for $\mathcal{N} > 100$. However, the number of populations used in the experiments considered are too small to consider that the large $\mathcal{N}$ asymptotic are applicable. To deal with smaller values of $\mathcal{N}$, we have thus constructed the empirical confidence intervals, as explained above in general.

In the next Section, we first illustrate how the empirical confidence interval computation algorithm is applied to calculate the confidence intervals for the estimator $\hat{\nu}$.

4.2.5 Empirical Confidence Interval of $\hat{\nu}$

For fixed values of s and $\mathcal{N}$, for each potential value $u \in GR$ of the unknown parameter ν, we first simulate 1100 trajectories of the evolutionary population process. We generate for each pair of (s, μ), corresponding empirical distributions of $T_{wait} = T_\beta - T_{bot}$, and the empirical distributions of T_β.

After this preliminary step, we use the empirical distributions of T_β repeatedly to generate 1000 random samples of $\mathcal{N}$ independent virtual values $(T_\beta^1, \ldots, T_\beta^{\mathcal{N}})$. To each one of these random samples of size $\mathcal{N}$, we apply the estimation algorithm

above to compute the corresponding "virtual" value of the estimator $\hat{\nu}$. The histograms of the 1000 virtual values of $\hat{\nu}$ thus provides the empirical distribution of the estimator $\hat{\nu}$, when the current fixed potential value u is the true value of the logarithmic mutation rate. From these, we can compute the quantiles η_- and η_+, for fixed confidence level α. Note that these depend on the true unknown value u and the other model parameters s and $\mathcal{N}$.

After repeating this process for each potential value $u \in GR$, we can then display the two curves $\eta_- : u \to \hat{\eta}_-(u)$ and $\eta_+ : u \to \hat{\eta}_+(u)$. Once these two curves are generated, they can then be used to generate a confidence interval at confidence level α for the unknown true value of ν, as soon as the value of the estimator $\hat{\nu}$ has been computed. For the asymptotic situation where the number of populations $\mathcal{N} \to \infty$, formula (4.12) shows that $\eta_- < \eta_+$ are continuous, non-decreasing functions. Thus,

$$P(\eta_-(\theta) \leq \theta \leq \eta_+(\theta)) = \alpha$$

linking the random estimator $\hat{\theta}$ and the true value θ of the parameter. Also,

$$\hat{\theta} \leq \eta_+(\theta) \quad \textit{and} \quad \eta_-(\theta) \leq \hat{\theta} \tag{4.13}$$

and by construction, this (equation (4.13)) is true 90% of the time. We can thus write an equivalent form of these inequalities, and can solve for $\eta_+(x) = y$ or $x = Z_+(\eta_+(x))$ and $\eta_-(x) = y$ or $x = Z_-(\eta_-(x))$, and thus inequalities (4.13) are equivalent to

$$Z_+(\hat{\theta}) \leq \theta \quad \textit{and} \quad \theta \leq Z_-(\hat{\theta}) \tag{4.14}$$

which gives $Z_+(\hat{\theta}) \leq \theta \leq Z_-(\hat{\theta})$ and $P(Z_+(\hat{\theta}) \leq \theta \leq Z_-(\hat{\theta})) = \alpha$.

The functions $A(\hat{\theta})$ and $B(\hat{\theta})$ are given as above in equation (4.2), and the confidence

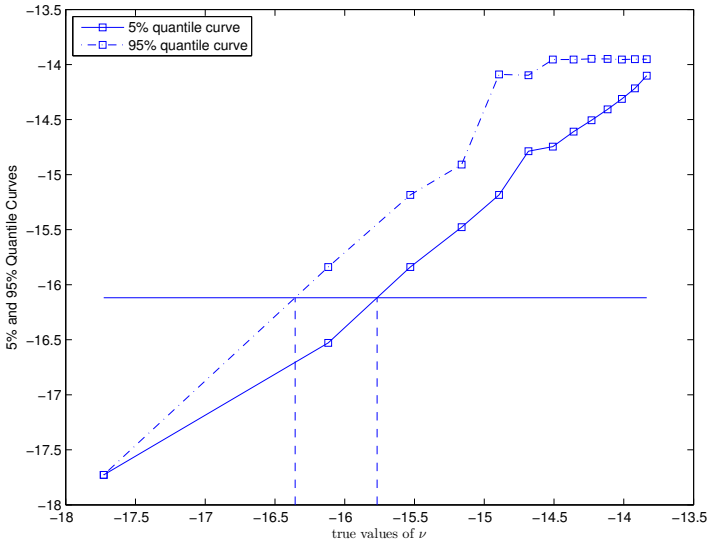


Figure 4.13: Confidence intervals for $\hat{\theta}$ computed using the quantile curves.

interval is then defined as $CI(\hat{\theta}) = [A(\hat{\theta}), B(\hat{\theta})]$ where $A(\hat{\theta})$ and $B(\hat{\theta})$ are the abscissas of the points having ordinate $\hat{\theta}$ on η_+ and η_- respectively.

Figure 4.13 displays an example of the quantile curves, and how the confidence interval is computed by applying the inversion algorithm explained above. This example displays the quantile curves η_+ and η_- for instance for $s = 0.1$ and $\mathcal{N} = 50$, and $\alpha = 90\%$. If the experimental data, for instance, were generated by 50 populations, for $s = 0.1$ and some unknown ν, let the estimated value to be suppose $\hat{\theta} = -16.12$, then figure 4.13 shows that the 90% confidence interval corresponding to the value $\hat{\theta} = -16.12$ is $CI(-16.12) = [-16.36, -15.77]$, since the horizontal line of ordinate -16.12 intersects the curves η_+ and η_- at points with abscissas -16.36 and -15.77. The interval $CI(-16.12)$ should thus contain the unknown true $\hat{\theta}$ with probability 90%.

For $\alpha = 90\%$, we have validated numerically that η_- and η_+ are nondecreasing and that $\eta_- < \eta_+$, holds for moderate numbers of populations $\mathcal{N}$, and that the preceding algorithm correctly generates families of confidence intervals. This numerical verification was performed for each $s \in \{0.01, 0.02, \ldots, 0.2\}$ and for each $\mathcal{N} = 11, 30, 50, 70, 90, 100$. However, the property that $u \in CI(u)$ may fail for extreme or isolated values of u. This point has led us to algorithmically improve the estimator $\hat{\nu}$ to decrease its bias, as indicated below.

4.2.6 Final Re-centered Estimator

Extensive numerical exploration shows that for moderate values of $\mathcal{N}$, the estimator $\hat{\nu}_{int}$ of ν tends to underestimate the true value of ν (figure 4.12 for example displays empirical histogram $\hat{\nu}_{int}$ when $\mathcal{N} = 11$), so that this slant naturally tends to "push" the estimator $\hat{\nu}_{int}$ towards the left boundary of the confidence interval $CI(\hat{\nu})$, and sometimes beyond it. Hence we have implemented and tested the following empirical re-centering technique for the estimator $\hat{\nu}_{int}$.

Fix the number of populations $\mathcal{N}$. Once the selective advantage s has been estimated, and hence can be considered known, we compute as indicated above, the histogram $Dis\ \hat{\nu}_{int}(s, u, \mathcal{N})$ of the estimator $\hat{\nu}_{int}$, for each potential value $u \in GR$ of the unknown true ν. Now for each fixed $u \in GR$, perform the following steps:

1. Apply the algorithm (4.2.5) to compute the empirical confidence interval for any potential value ϕ of the estimator $\hat{\nu}_{int}$. We generate the 90% confidence intervals $CI(\phi_k) = [A(\phi_k), B(\phi_k)]$ for each potential value ϕ_k for $k = 1, \ldots, 1000$.

2. By repeated independent sampling of the histogram $Dis\ \hat{\nu}_{int}(s,u,\mathcal{N})$, generate 1000 virtual values ϕ_k of the estimator $\hat{\nu}_{int}$, and compute the corresponding 1000 intervals $CI(\phi_k)$,

3. Use these intervals to generate 1000 random coefficients

$$coeff_k = (u - A(\phi_k))/(B(\phi_k) - A(\phi_k))$$

4. Compute the empirical median $\gamma(u)$ of the 1000 coefficients $coeff_k$.

We have thus pre-computed a deterministic numerical function $\gamma(u)$ for each $u \in GR$. For each $\phi \in GR$, let $\Gamma(\phi)$ be the average of $\gamma(y)$ over all $y \in CI(\phi) \cap GR$, and define the pondering coefficient $\delta(\phi) \in [0,1]$ by

$$\delta(s) = \begin{cases} \Gamma(\phi) & \text{if } 0 \leq \Gamma(\phi) \leq 1 \\ 0 & \text{if } \Gamma(\phi) < 0 \\ 1 & \text{if } \Gamma(\phi) > 1 \end{cases} .$$

When we use the results of the experiment to compute the associated actual value ϕ of the estimator $\hat{\nu}_{int}$, we may assume with probability larger than 90%, that the unknown true parameter ν lies somewhere in the interval $CI(\phi) = [A(\phi), B(\phi)]$. To get a better idea of how the unknown ν is positioned in $CI(\phi)$, it would be quite useful to know the pondering coefficient $\gamma(\nu)$ but this is not feasible since we do not know ν. Hence we replace this unreachable coefficient $\gamma(\nu)$ by its quite computable conditional average value $\delta(\phi)$ given the observed value ϕ of $\hat{\nu}_{int}$. A fairly natural guess for the unknown ν would be the barycenter $(1-\gamma(\nu))A(\phi)+\gamma(\nu)B(\phi)$, so we conclude that a good computable estimation of ν should be given by $(1-\delta(\phi))A(\phi)+\delta(\phi)B(\phi)$. By

Table 4.2: Displays the pre-computed three deterministic functions $A(\phi)$, $B(\phi)$, and $\delta(\phi)$ for all $\phi \in GR$.

ϕ	$A(\phi)$	$B(\phi)$	$\delta(\phi)$
-16.12	-17.72	-15.42	0.0006
-15.53	-17.72	-14.68	0.0939
-15.56	-17.72	-14.74	0.0939
-14.89	-15.81	-14.01	0.4638
-14.68	-15.69	-13.89	0.5515
-14.51	-15.53	-13.81	0.6499
-14.36	-15.34	-13.81	0.7219
-14.23	-15.20	-13.81	0.7219
-14.12	-15.16	-13.81	0.7219
-14.01	-14.91	-13.81	0.7219
-13.92	-14.50	-13.81	0.8386
-13.84	-14.47	-13.81	0.8386

simulations described above, we can pre-compute the three deterministic functions $A(\phi), B(\phi)$, and $\delta(\phi)$ for all ϕ in the grid GR. These are displayed below in table 4.2.

Given the $\mathcal{N}$ observed trajectories recorded in an actual experiments, we can first compute the intermediary estimator $\hat{\nu}_{int}$ of the unknown true ν. We then define and compute a new final estimator of ν, denoted $\hat{\nu}$ by

$$\hat{\nu} = (1 - \delta(\hat{\nu}_{int}))A(\hat{\nu}_{int}) + \delta(\hat{\nu}_{int})B(\hat{\nu}_{int}).$$

This construction implements a self adaptive re-centering of the intermediary estimator, so one should expect the final estimator $\hat{\nu}$ to have less bias and more accuracy than the intermediary estimator $\hat{\nu}_{int}$. Figure 4.14 displays an example of the empirical histogram of the final estimator $\hat{\nu}$ of $\nu = \log(\mu)$, when $s = 0.10$ is fixed and for $\mathcal{N} = 11$.

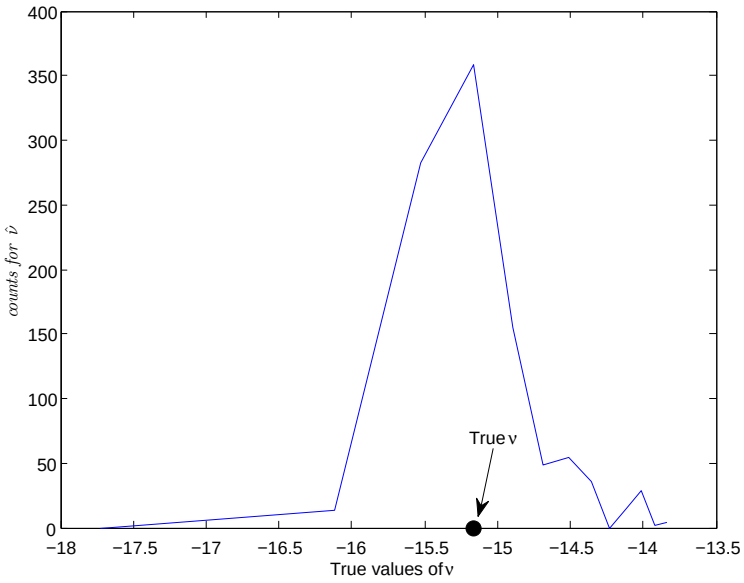


Figure 4.14: The empirical histogram of $\hat{\nu}$ for fixed $s = 0.12$ and $\mathcal{N} = 11$. The true value of $\nu = -15.16$. We see that re-centering improves the intermediary estimator $\hat{\nu}_{int}$ (figure 4.12.)

4.2.7 Accuracy of the Logarithmic Mutation Rate Estimator

We have studied and compared the performances of the two estimators $\hat{\nu}_{int}$ and $\hat{\nu}$ for the representative values of s and moderate values of $\mathcal{N}$. The estimator $\hat{\nu}$ clearly performs better than $\hat{\nu}_{int}$ over the whole range of potential values of $\nu = \log \mu$. We now evaluate the quantitative performance of the estimator $\hat{\nu}$. For each fixed triplet $s, \nu, \mathcal{N}$, we calculate the empirical histogram of $\hat{\nu}$. We use the histogram $Hist\ \hat{\nu}_{int}(s, u, \mathcal{N})$ to generate 1000 values of the intermediary estimator $\hat{\nu}_{int}$. Then, applying the self adaptive re-centering algorithm as above, we obtain 1000 values of the final estimator $\hat{\nu}$.

We can then compute the three indicators described above in Section 4.1, the $Bias$, the error of estimation Err, and the average width $l(CI)$ of the 90% confidence

intervals for the final estimator $\hat{\nu}$. Since, the final or intermediary estimator of $log\ \mu$ is computed based on a given s, we present the performances of the final estimator $\hat{\nu}$ for a single representative value of $s = 0.1$ and for 4 different values of $\mathcal{N} = 11, 30, 50, 100$. For each of these cases, we display the curves for the $Bias$, Err, and $l(CI)$ of $\hat{\nu}$ as a function of true ν as displayed in figure 4.15.

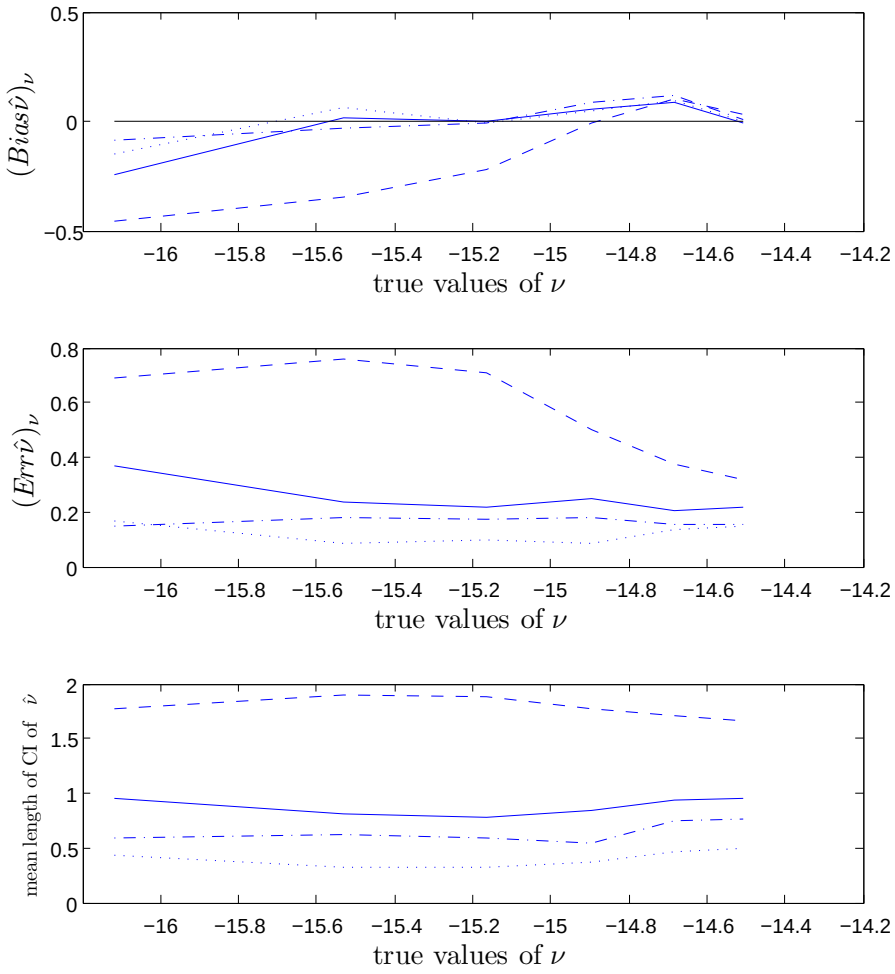


Figure 4.15: Accuracy of the final estimator $\hat{\nu}$ of the logarithmic mutation rate : the bias and the error of estimation, and average width of confidence interval are displayed as functions of $\nu = \log \mu$ for $s = 0.1$ and $\mathcal{N} = 11, 30, 50, 100$. The dotted line curves correspond to $\mathcal{N} = 11$, the solid line curves correspond to $\mathcal{N} = 30$, the dotted and dot line curves correspond to $N = 50$, and the tiny dotted curve correspond to $\mathcal{N} = 100$.

We see sufficient improvement in the accuracy when $\mathcal{N}$ increases from 11 to 30. For $\mathcal{N} = 30, 50, 100$, the *Bias* and the average error *Err* of estimation of the final estimator $\hat{\nu}$ are both inferior to 3% of the true value of ν. The 90% confidence intervals generated by this estimator have an average width inferior to 10% of the true value of ν. The *Bias* and the average error *Err* of the estimator $\hat{\nu}$ are both small even for the moderate number of populations $\mathcal{N} = 30$, so increasing this number to 100 does not reduce the two quantiles significantly. However, the average width of the 90% confidence intervals for ν decreases significantly when $\mathcal{N}$ increases from 30 to 100, and is in fact roughly multiplied by the theoretical factor $\sqrt{30/100} = 0.54$. When the number $\mathcal{N}$ of the experimental populations becomes large, one can show that our final estimator $\hat{\nu}$ of the logarithmic mutation rate becomes approximately Gaussian, with an asymptotic error of estimation of the order of

$$Err \approx \frac{P_{bot}^{3/2}}{(1 - P_{bot})\sqrt{\mathcal{N}}}.$$

For the TC experiments, these asymptotic errors of estimation on $\nu = \log \mu$ have a width inferior to $2/\sqrt{\mathcal{N}}$, but this will only become effectively valid for $\mathcal{N} > 100$. Table 4.3 below displays example of how $\frac{P_{bot}^{3/2}}{1-P_{bot}}$ varies with different s and μ.

Table 4.3: Table displaying the values for $\frac{P_{bot}^{3/2}}{1-P_{bot}}$ for different s and μ.

$s \backslash \mu$	2×10^{-7}	5×10^{-7}
0.10	0.014	0.056
0.15	0.016	0.066

4.2.8 Another Re-centering of ν and its Accuracy

We introduce another re-centering algorithm to re-center the preliminary estimate $\hat{\nu}_{int}$ of $\nu = \log \mu$ by computing the new re-centered estimator by taking the average of the 90% confidence interval as computed for the preliminary estimator, by using the algorithm as described in Section 4.2.5. For a fixed number of populations $\mathcal{N}$ and the selective advantage s, we compute as above the empirical histogram $Dis\hat{\nu}_{int}(s, \mu, \mathcal{N})$ of the estimator $\hat{\nu}_{int}$ for each potential value $u \in GR$ of the unknown true ν. Now, for a fixed $u \in GR$, we apply algorithm 4.2.5 to compute the empirical confidence interval for any potential value ϕ of the estimator $\hat{\nu}_{int}$. This generates 90% confidence intervals $CI(\phi_k) = [A(\phi_k), B(\phi_k)]$ for each potential value ϕ_k for $k = 1, \ldots, 1000$. By repeated independent sampling of the histogram $Dis\,\hat{\nu}_{int}(s, \nu, \mathcal{N})$, generate 1000 virtual values ϕ_k of the estimator $\hat{\nu}_{int}$, and compute the corresponding 1000 intervals $CI(\phi_k)$. We use these 1000 intervals $CI(\phi_k)$ to compute the re-centered estimator $\hat{\nu}$ as follows:

$$\hat{\nu}(\phi_k) = \frac{A(\phi_k) + B(\phi_k)}{2}$$

Thus generating empirical distributions of the re-centered estimator $\hat{\nu}$ of ν by considering the average of the 90% confidence intervals $CI(\phi_k)$ for each fixed $u \in GR$. Figure 4.16 displays example of the empirical histogram for the re-centered estimator when the re-centering is performed as above by considering the average of the confidence intervals for each potential ϕ_k for $k = 1, \ldots, 1000$. We studied the accuracy of this re-centered estimator $\hat{\nu}$ of ν and in figure 4.17 we display the results for the mean length of the confidence intervals, as the size of the number of wells $\mathcal{N}$ increases from 11 to 50, for a fixed value of s. We see that the mean length of confidence intervals

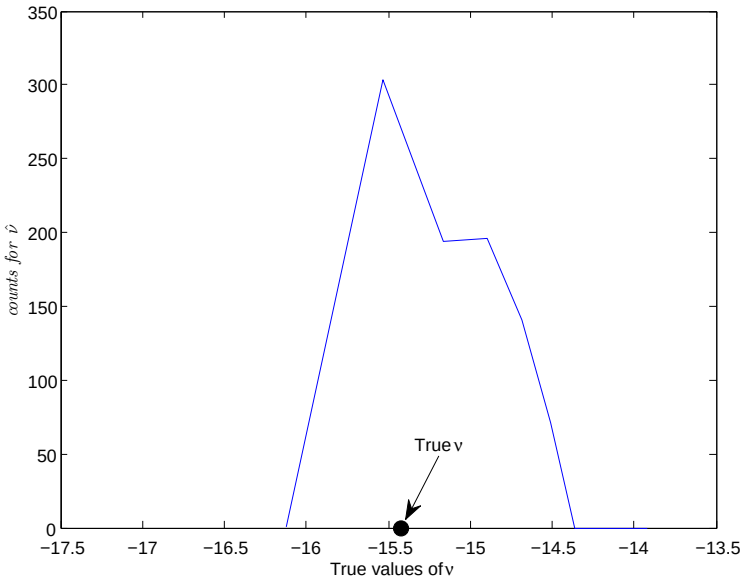


Figure 4.16: This figure displays the empirical distribution of re-centered $\hat{\nu}$ using the average of the confidence intervals, for a fixed value of $s = 0.10$ and for $\mu = 2 \times 10^{-7}$, and $\mathcal{N} = 11$.

decreases as the size for the number of wells $\mathcal{N}$ increases. Later in chapter 5, we will show a comparison of the estimator $\hat{\nu}$ (re-centered using the average of confidence intervals) with the estimator of ν after complementary sub-sampling, and how the accuracy is affected with different sub-sampling sizes (Section 5.6.1).

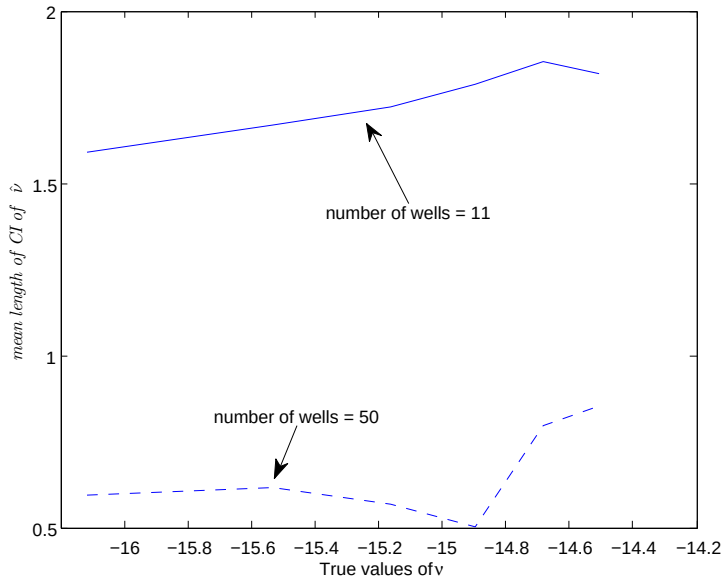


Figure 4.17: This figure displays the mean length of the confidence intervals for true values of ν, for a fixed $s = 0.10$, and for different $\mathcal{N} = 11, 50$. The mean length of the confidence intervals decreases as the number of wells $\mathcal{N}$ increases.

CHAPTER 5

Application to Experimental Data and Effect of Complementary Sub-Sampling

In the above chapters, we have introduced and studied a detailed Poisson process model for the first step evolutionary dynamics of an asexual bacterial population evolving under the simplifying assumption that a single type of irreversible mutation is available to the population. The model formalizes experiments that evolved $\mathcal{N}$ populations starting with individuals having, except for a neutral marker, identical genotypes, where each initial population undergoes a series of daily {growth+dilution} cycles. The simulation study mentioned above predicts that the estimator $\hat{s}$ of the

selective advantage s has quite good accuracy within the realistic ranges of the model parameters for the TC experiments, for estimating the selective advantages of the first adaptive mutation that occurs and influences evolutionary dynamics. We will test this prediction by applying our estimation algorithms as explained above to the experimental data collected in the TC experiment and compare the predicted and experimentally measured estimates.

5.1 Analysis of the Experimental Data

Denote by $Pop_1, \ldots, Pop_k$ the k observed *E.coli* populations. As explained above (in chapter 3), every day, after the growth saturation of these k populations, a sample of very large fixed size N_0 is randomly sampled, by dilution, from each population Pop_j, and transferred to a fresh well. The frequencies w_t and r_t of the two cell marker types in a fresh well on day t are then assayed from a complementary sub-sample randomly extracted from the newly transferred population.

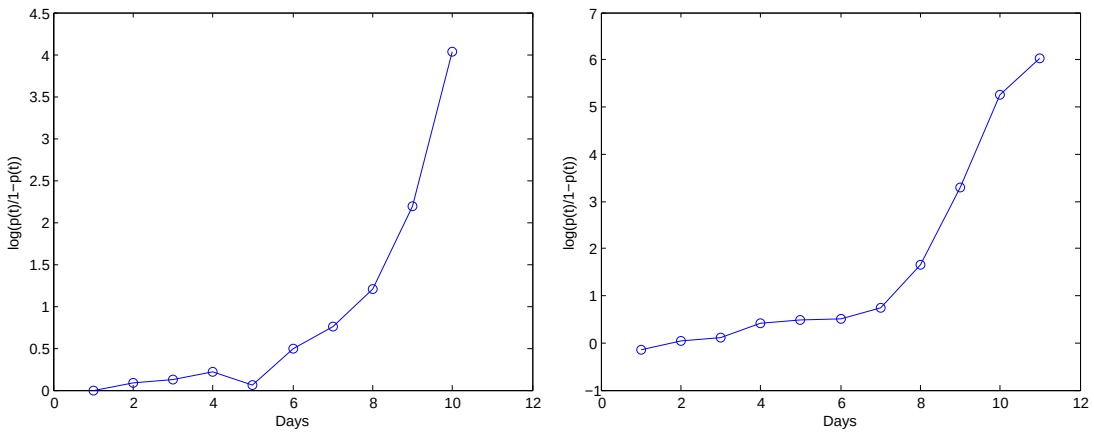
For TC experiments, the size N_{sub} of the complementary sub-sampling is moderate, typically ranging between 300 and 400. A sample size of 300 is still reasonable to estimate frequencies, but the experimental value for $p(t)$ will have associated sampling error which cannot be a priori neglected in the accuracy study of the parameter estimators. Hence application of our estimator $\hat{s}$ to actual TC experimental data requires several systematic numerical modifications to remain efficient. We will give an analysis of the complementary sub-sampling effect in the later part of this chapter.

In table 5.1, we present an example of a typical data set for the TC experiments we

have studied. This example displays the recorded instances of daily red and white cells counts for one replicate population in the first TC experiments, which start with a pure population of Ara-1 ancestor cells. Figure 5.1 displays two examples of the curve $\log \frac{p(t)}{1-p(t)}$ where $p(t)$ is the frequency of the white winning marker. This figure displays examples for population 2 and population 10 obtained from the TC experiments with pure population of ancestor cells in the beginning of the experiments.

Table 5.1: Example of recorded daily red and white cell counts data for Pop_1 experimental population of *E. coli* with initially identical ancestor genotype.

days	1	8	15	23	25	29	32	34	36	50
# Red	26	138	115	42	172	425	320	300	300	250
# White	24	136	75	22	48	18	1	0	0	0



(a) (b)

Figure 5.1: Curve $\log \frac{p(t)}{1-p(t)}$ displayed over days for population 2 in figure (a) and for population 10 in figure (b), for the 1st experimental data of the TC experiments; which consisted of the pure ancestor population at the beginning of the experiments. the dots represents the days at which the counts for red and white were recorded for the experimental data.

5.2 Comparison of Estimated and Actual Selective Advantages

Here we compare the estimated and the actual selective advantages for one TC experiment. Out of the 6 distinct TC experiments, the "ground truth values" are only available for the first TC experiment, where the initial genotype is the ancestor genotype (with no mutation present in the beginning). We focus on the $\mathcal{N} = 11$ independently evolved and monitored populations of initially identical ancestor cells. We present the analysis for 10 of the 11 independently evolved and monitored populations of initially identical ancestor cells, as described above, excluding one population which exhibited unusual complex marker dynamics that are likely to indicate the presence of more than one beneficial mutation. These populations are monitored until an emerging mutant reaches fixation.

To evaluate the "ground truth values" of s, complementary experimental manipulations are needed as explained above in Section 2.2. For each population, four individuals of the winning marker type were isolated from a time point as close as possible to when the winning marker reached a frequency of greater than 0.9. These individuals are likely to differ from the original ancestral genotype by the addition of the beneficial mutation that drove the change in the marker dynamics. The fitness of each individual was measured relative to the ancestor (Lenski et al. (1991) [35]), by using a GFP expressing derivative of the ancestor as the reference. The use of GFP marker enabled distinguishing competing ancestor and evolved clone sub-populations by flow cytometry (as explained in Section 2.2). The mean of these measures provides

a direct experimental value $s_{dir}(j)$ for the actual selective advantage of the beneficial mutation causing the winning marker to fix in population Pop_j. These values for s_{dir} are called the "ground truth values" of s.

For each population, we also computed an estimate $\hat{s}(j)$ of the selective advantage by applying our algorithmic estimator to the sequence of recorded experimental estimates of the frequency $p(t)$. The 10 pairs of $s_{dir}(j)$ and $\hat{s}(j)$ together with their errors (standard deviations) are displayed in the figure 5.2.

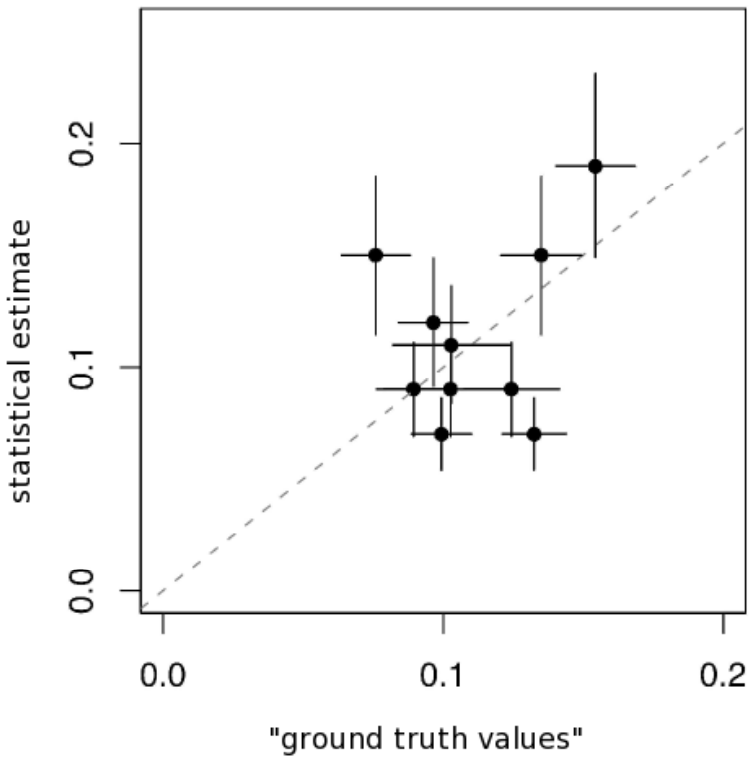


Figure 5.2: Comparison of "ground truth values" to our estimates of selective advantages in 10 evolved populations. The experimental "ground truth values" s_{dir}, and our estimated values $\hat{s}$ are displayed. The horizontal line indicates 95% confidence intervals on s_{dir} and the vertical lines are $\pm$ standard deviations for the estimation errors attached to $\hat{s}$.

The mean 0.111 of the 10 experimental "ground truth values" $s_{dir}(1), \ldots, s_{dir}(10)$

is quite close to the mean 0.114 of our estimates $\hat{s}(1), \ldots, \hat{s}(10)$. Moreover, the square root of the mean quadratic differences (MQD) between our estimated values and the "ground truth values" of the selective advantages in these 10 populations is 0.036. The standard deviation of the "ground truth values" $s_{dir}(j)$ has been evaluated to be of the order of $\sigma_{dir} = 0.007$. The standard deviation of the estimation errors for our estimators $\hat{s}(j)$, obtained by analyzing the observed experimental data is close to $\hat{\sigma} = 0.03$ for each of the 10 populations studied.

Let $\sigma^2_{dir}(j)$ be the variance corresponding to $s_{dir}(j)$ and let $\hat{\sigma}^2(j)$ be the variances corresponding to $\hat{s}(j)$. Then we want to compare the two variances $\sigma^2_{dir}(j)$ and $\hat{\sigma}^2(j)$. Consider $\sigma^2_{dir}(j)$ to be known, then we compare to $\hat{\sigma}^2(j)$. Let $V_1 = \sigma^2_{dir}(j) + \hat{\sigma}^2(j)$ be the combined variance of the two sets of estimates, and let $V_2 = \sum_{i=1}^{10}(\hat{s}(i) - s_{dir}(i))^2$ be the sum of squared differences between direct and model estimates. To further test the agreement of the direct and model-based estimates, we calculated the statistic

$$X^2 = \frac{V_2}{V_1}.$$

Under our null hypothesis, each difference $s_{dir}(j) - \hat{s}(j)$ follows a distribution with mean zero and variance V_1, and X^2 is expected to follow a χ^2 distribution with 10 degrees of freedom. Thus performing the hypothesis test, we test the null hypothesis that these two variances are equal.

Hence we find that our model estimated values of selective advantages are compatible with the "ground truth values".

5.3 Estimate of Beneficial Mutation Rate

Direct measurement of occurrence rates for beneficial mutations is notoriously difficult, and is currently possible for only a limited number of strain-mutation combinations [8]. Thus, standard biological experiments cannot directly assess the accuracy of our estimates for the occurrence of beneficial mutations as derived from the experimental data.

The overall estimate of ν, as computed from the model, for the 10 replicate populations depend on the estimate of s. If we take the mean of the 10 estimates $\hat{s}$ as an overall estimate of s, the estimate of logarithm of mutation rate using the method described above in (4), for the simulation model is given by $\hat{\nu} = -15.16$ with error size $|\hat{\nu} - \nu|$ is 0.7. This value corresponds to a mutation rate μ between 1.3×10^{-7} and 5.2×10^{-7}. The square root of the mean quadratic error $\sqrt{mean(\hat{\nu} - \nu)^2}$ is 0.8.

5.4 Effect of the Complementary Sub-sampling

In the experimental growth evolution experiments, as explained above, after the growth saturation of the cells, every day, first a sample of a large fixed size N_0 is randomly sampled, by dilution and transferred to a fresh well. A complementary sub-sample, of size N_{sub} is then randomly extracted from this newly transferred population, and allowed to grow on culture plates, to extract current frequency of color markers by visual counts. The maximum of number of cells, after growth on culture plates is restricted to between 300 and 400. The frequencies for the two

color marker types cells are then assayed at different days from these culture plates. For the TC experiments, the size of this sub-sampling N_{sub} is moderate, typically between 300 and 400.

In our simulations of the preceding process, at the end of each day, after the "growth + first dilution" cycle has been simulated, the complementary random sub-sample of size N_{sub} is virtually extracted. This sub-sampling follows a binomial distribution $\sim Bin(N_{sub}, pr)$, where the parameters for this binomial are the total sub-sampling size given by N_{sub} and probability pr denotes the proportion of cells for any one marker type in the population. The sub-sampling generated random number of cells for the white marker type, for each day t.

5.4.1 Algorithm for Automatic Extraction of the "Almost Linear Growth" Time Segment

Let $g(t) = \log \frac{p(t)}{(1-p(t))}$. We present an algorithm for the automatic extraction of the fast almost linear increase " time segment" from the curves of $g(t)$, to analyze the markers frequency data after the complementary sub-sampling.

As mentioned in Section 4.1.2, we have computed a simulation data base giving the daily numbers of red and white cells as counted *after* the daily complementary sub-sampling. For each pair of (s, μ) in their respective ranges, we have simulated 1000 such trajectories. We have developed an algorithm to automatically detect the first time the deviation from the straight line occurs in the trajectory $g(t)$, and extract, from the $g(t)$ trajectory, the time segment of roughly linear growth from this point

onwards. Since the complementary sub-sampling size is much smaller, we have a larger error attached to the $p(t)$, which increases the fluctuations in the observed trajectory of $g(t)$. The following algorithm allows us to extract the roughly linear time segment from the $g(t)$-trajectory .

1. For each day t starting from day 2 and going up to day T, break up the trajectory into two parts, the first, denoted by $[0, t]$ from day 1 to day t, and a part $[t, T]$, from day t to day T, where T is the first time at which the frequency of winner $p(t)$ reaches 99%.

2. Compute the best fit least squares line, called $f_t(u)$, modeling the $g(u)$-trajectory for u in $[0, t]$. The squared error of fit between $f_t(u)$ and $g(u)$ has mean $m(L, t)$, median $= med(L, t)$ and 75% quantile $Q(L, t)$. Similarly, compute the best fit least squares line f_T modeling the trajectory on $[t, T]$. The squared error of fit between $f_T(u)$ and $g(u)$ have mean $m(R, t)$, median $= med(R, t)$ and 75% quantile $Q(R, t)$. Figures 5.3, 5.4, and 5.5 displays example of these curves for a particular trajectory $g(t)$ displayed in figure 5.7.

3. Compute $\sigma_1(t) = m(L, t) + m(R, t)$, $\sigma_2(t) = med(L, t) + med(R, t)$ and $\sigma_3(t) = Q(L, t) + Q(R, t)$, where $\sigma_i(t)$ denotes the sum of the mean ($i = 1$), median ($i = 2$), or the 75th quantile ($i = 3$) for the best fit least squares line f_t and f_T for each day t. Figures 5.3, 5.4, and 5.5 displays these curves.

4. Compute $D_0 =$ 't' which minimizes $[\min\{\sigma_1(t), \sigma_2(t), \sigma_3(t)\}]$. D_0 is the the day on which the first significant trajectory deviation occurs, and is a candidate for T_{begin}. However, it is possible that D_0 overestimates T_{begin} by a few days. We

will correct for this error later. Figure 5.6 displays the three sums computed above, $\sigma_1(t)$, $\sigma_2(t)$, and $\sigma_3(t)$.

5. Let $\hat{D} = \max(D_0, \text{argmin}_t \sigma_1(t))$. $\hat{D}$ is either D_0, or is the day on which the minimum sum of mean squared errors is obtained. Starting from $\hat{D}$, check to see if, for each $1 \leq i \leq T - \hat{D}$

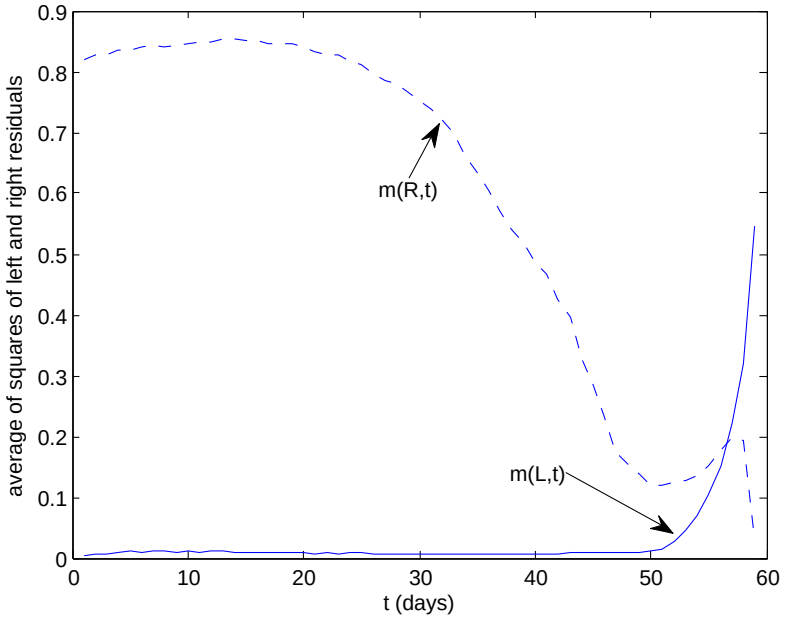
$$\sigma_1(\hat{D} + i) \geq \sigma_1(\hat{D} + i - 1).$$

If i_0 is the first i for which $\sigma_1(\hat{D} + i_0) < \sigma_1(\hat{D} + i_0 - 1)$, we let $D_1 = \hat{D} + i_0$. If this condition is never met, set $D_1 = T_{fix}$. D_1 is our preliminary estimate for T_{end}, but it is possible that D_1 underestimates T_{end} by a few days.

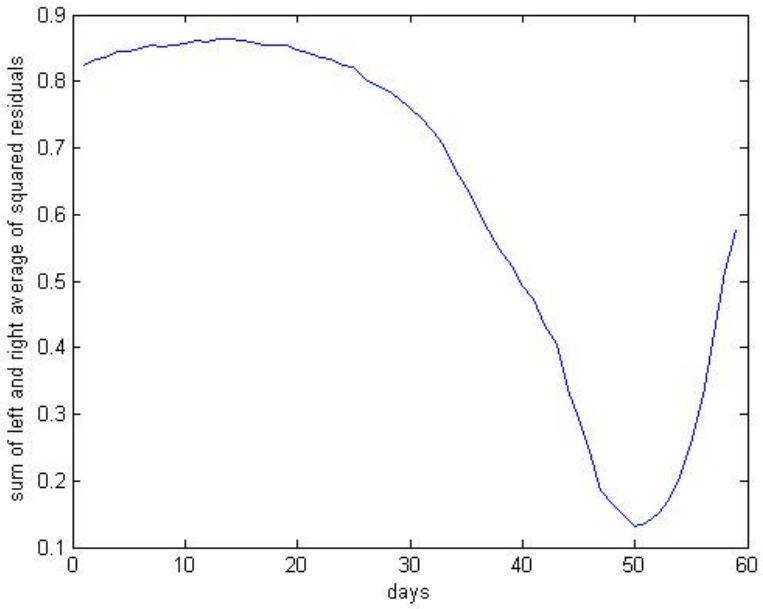
6. We now adjust the preliminary estimates D_0 and D_1 to get our estimates for T_{begin} and T_{end}. Fix a length parameter $L = 5$, and consider the intervals $[D_0 - L, D_0 + L]$ and $[D_1 - e, D_1 + e]$. These intervals should be truncated so that no part of these intervals lies outside $[0, T_{fix}]$. Pick a day $d_1 \in [D_0 - e, D_0 + e]$ and $d_2 \in [D_1 - e, D_1 + e]$. The trajectory now has three parts: $[0, d_1]$, $[d_1, d_2]$ and $[d_2, T]$. Compute the best fit least squares line for these three parts. On each part, compute the average squared error of the residuals, and sum over the three parts to obtain $\sigma(d_1, d_2)$. Then

$$(T_{begin}, T_{end}) = \text{argmin}_{d_1, d_2} \sigma(d_1, d_2).$$

Figure 5.7 displays the curve $g(t)$ along with the T_{begin} and T_{end} as obtained using the algorithm described above. Figure 5.8 displays another example for the curve $g(t)$ and plots the T_{begin} and T_{end} as computed by using the algorithm detailed above.

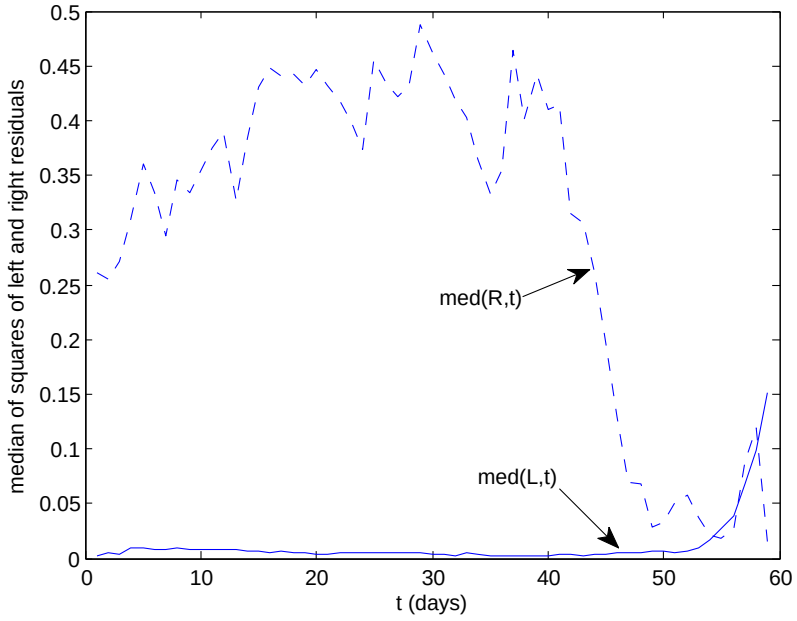


(a)

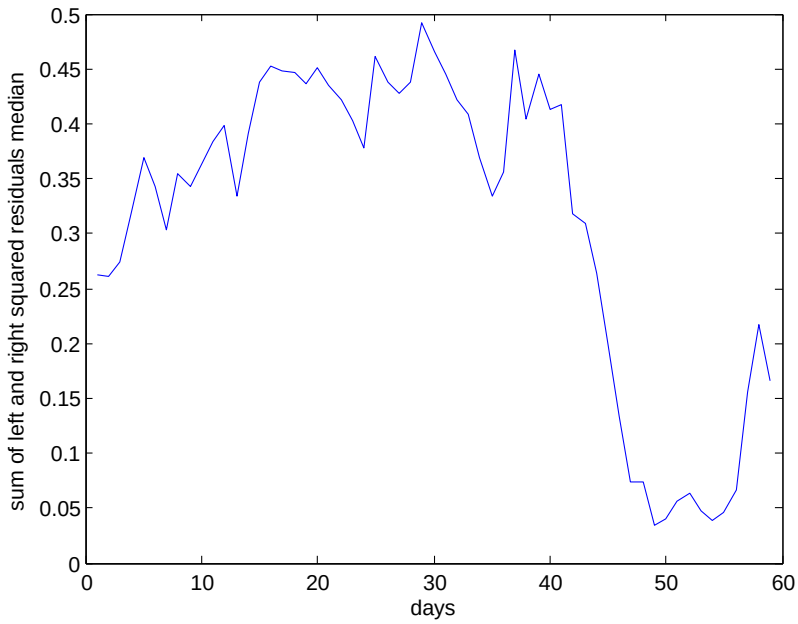


(b)

Figure 5.3: (a): Displays the plot for the average of the left ($m(L,t)$) and right ($m(R,t)$) residuals squares at each time point. (b): Displays the sum $\sigma_1(t)$.

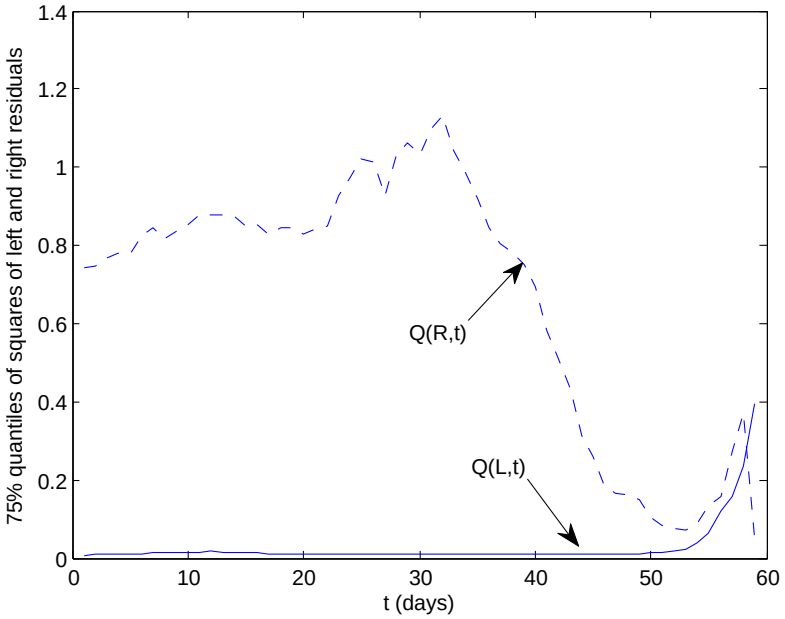


(a)

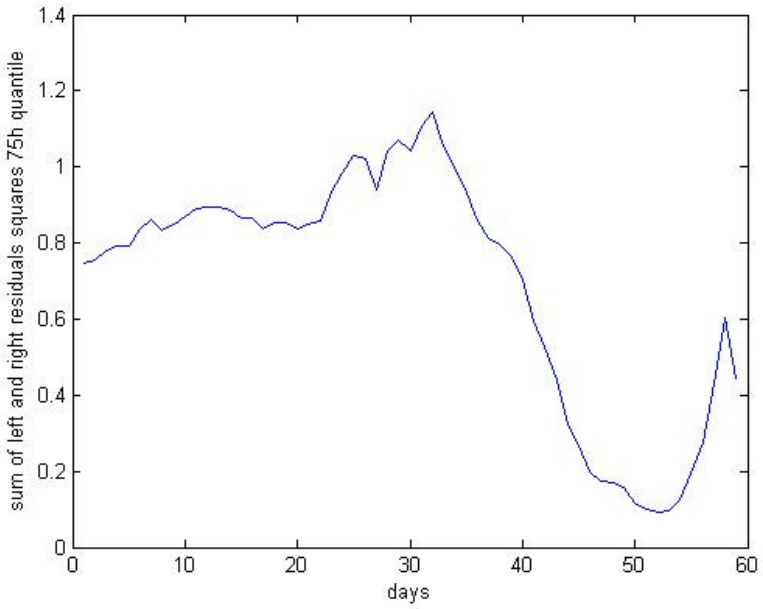


(b)

Figure 5.4: (a): Displays the plot for the median of the left $(m(L,t))$ and right $(m(R,t))$ residuals squares at each time point. (b): Displays the sum $\sigma_2(t)$.



(a)



(b)

Figure 5.5: (a): Displays the plot for the 75% quantile of the left ($Q(L,t)$) and right ($Q(R,t)$) residuals squares at each time point. (b): Displays the sum $\sigma_3(t)$.

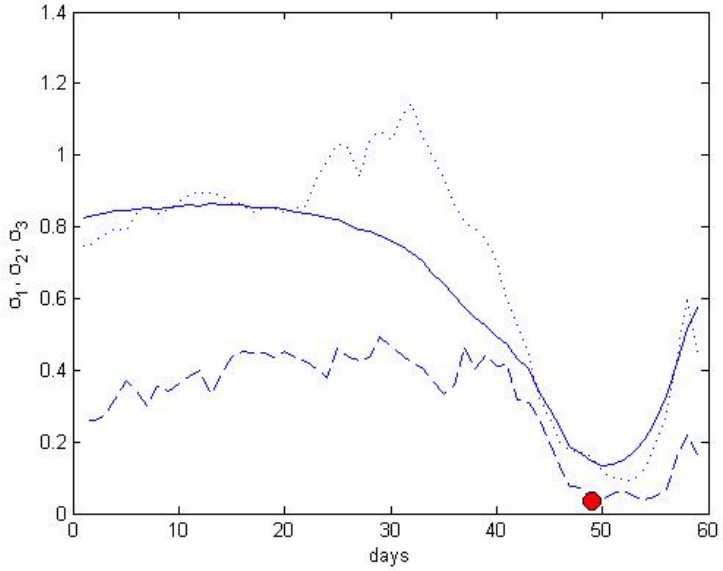


Figure 5.6: Plots for $\sigma_1(t)$ in the solid line, $\sigma_2(t)$ in the dashed line, and $\sigma_3(t)$ in the dotted line. The red round dot displays the "t" which minimizes $\min\{\sigma_1(t), \sigma_2(t), \sigma_3(t)\}$, which occurs at day 49.

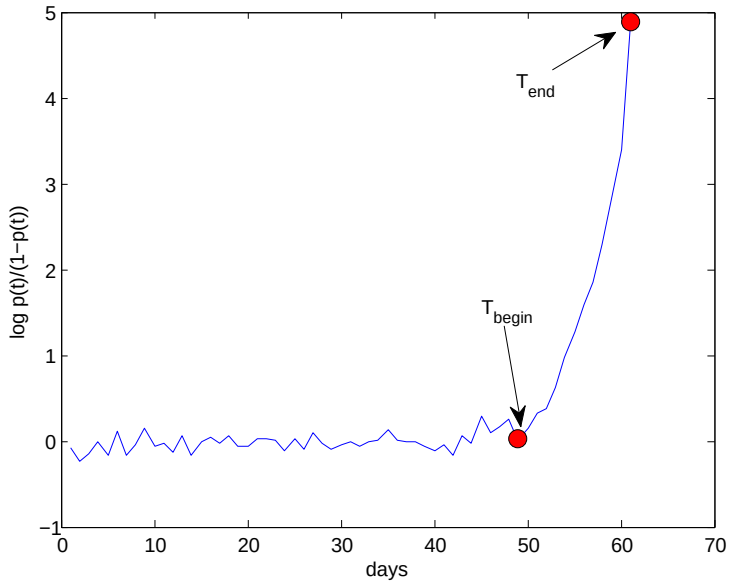


Figure 5.7: The curve $g(t)$ versus days, displaying the T_{begin} and T_{end} as computed using the algorithm above.

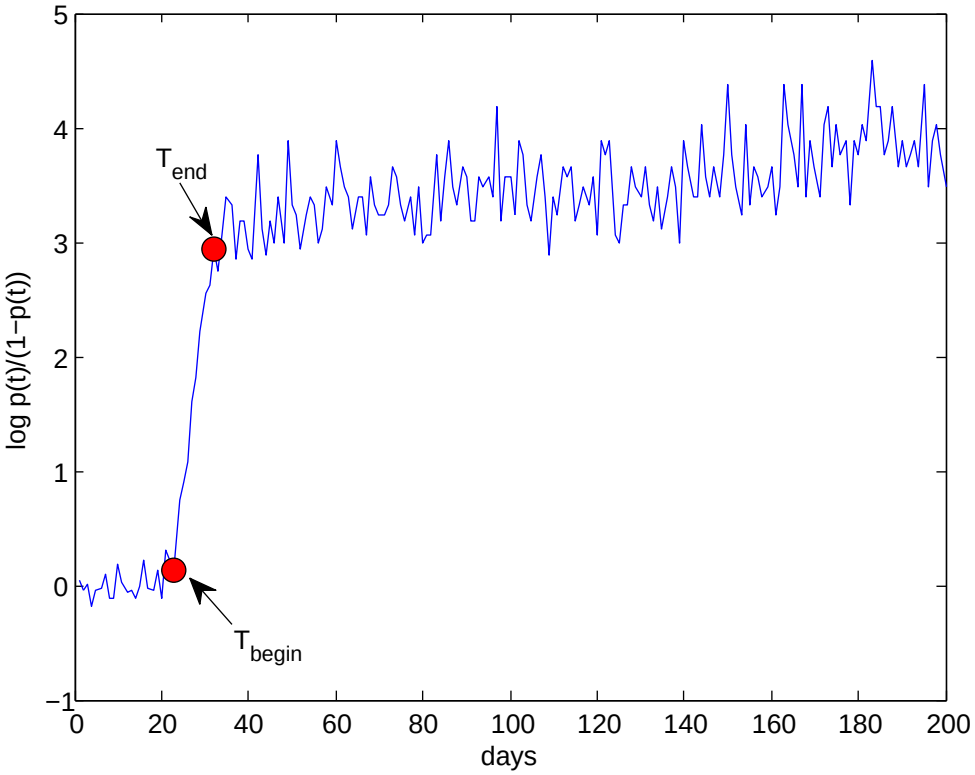


Figure 5.8: The curve $g(t)$ versus days, displaying the T_{begin} and T_{end} as computed using the algorithm above.

5.5 Estimation of s after Complementary Sub-sampling

Define a grid of one thousand (s, μ) values $GRID$ as the discretized rectangle built by pairing 20 values of $s \in [0.01, 0.2]$ with 50 values of $\mu \in [2 \times 10^{-8}, 10^{-6}]$.

For any given pair (s, μ) in GRID, we simulate a large number of trajectories of the evolution model parametrized by (s, μ). To each simulated $g(t)$-trajectory, we apply the automatic algorithm presented in Section 5.4.1 to extract the roughly linear growth time segment $SEG = [T_{begin}, T_{end}]$ of the $g(t)$-trajectory.

As in Section 4.2.1, we apply linear regression between T_{begin} and T_{end} to approximate $g(t)$ on SEG, and denote by $\hat{a}$ the estimated slope of this linear regression line. From equation (3.16), we see that the curve $g(t) = \log \frac{p(t)}{1-p(t)}$ is approximately linear in t with slope given by $\frac{s}{1+s} \log D$. Thus we have

$$\begin{aligned} \hat{a} &= \frac{s}{1+s} \log D \\ \hat{a}(1+s) &= s \log D \\ s(\log D - \hat{a}) &= \hat{a} \end{aligned}$$

and we generate our preliminary estimate of s as follows

$$\hat{s}_{pr} = \frac{\hat{a}}{\log D - \hat{a}}$$

Repeating this for all trajectories in the simulation data base, we obtain, for each pair (s, μ) in $GRID$, the associated empirical distributions $Dis\ \hat{a}(s, \mu)$ of $\hat{a}$ and this enables us to evaluate the empirical mean as well as the median $G(s, \mu)$ of the preliminary estimator $\hat{s}_{pr}$. Our intensive simulations definitely show that the median $G(s, \mu)$, is practically almost independent of μ, and is quite close to a linear function $G(s)$ of s, which essentially does not depend on μ, as displayed in figure 5.9.

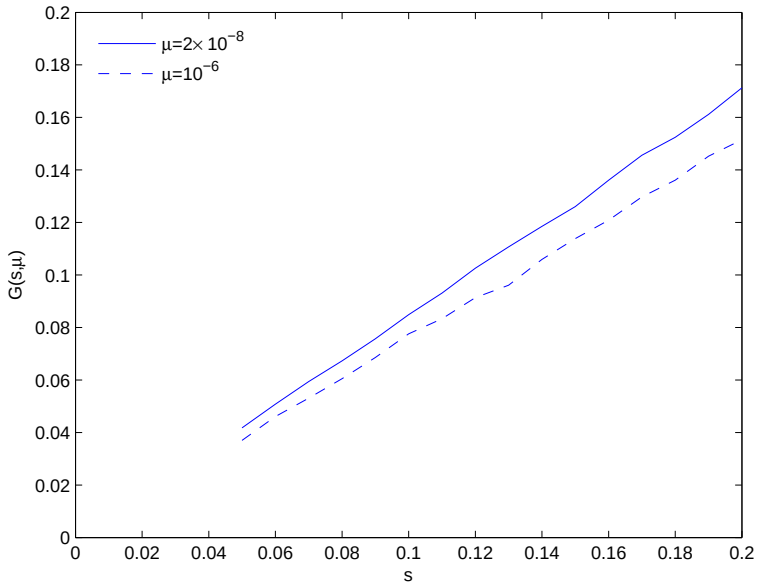


Figure 5.9: The empirical median $G(s,\mu) = G(s)$ of the preliminary estimator $\hat{s}_{pr}$ as a function of s for two extreme values of μ, is almost independent of μ, and is not equal to s.

A linear regression of $G(s,\mu)$ with respect to s generates the linear approximation

$$G(s) \simeq 0.8558s - 0.0009.$$

which is valid for all μ in GRID. The estimator $\hat{s}_{pr}$ is biased because the derivative $G(s,\mu)$ with respect to s is close to $0.8558 < 1$. We hence generate the new unbiased estimator of s by inverting the linear approximation of $G(s)$ as follows:

$$\begin{aligned} \hat{s} &= (\hat{s}_{pr} + 0.0009)/0.8558 \\ &= \left(\frac{\hat{a}}{\log D - \hat{a}} + 0.0009\right)/0.8558 \\ &= \left(\frac{\hat{a}}{5.3 - \hat{a}} + 0.0009\right)/0.8558 \\ &= \frac{1.17\,\hat{a} + 0.005}{5.29 - \hat{a}} \end{aligned} \tag{5.1}$$

Since $G(s) = G(s,\mu)$ is an increasing function of s, the median of the estimator $\hat{s}$ is

$G^{-1}(\cdot)$ evaluated at the median of $\hat{s}_{pr}$, $G(s)$. So the median of $\hat{s}$ is $G^{-1}[G(s)] = s$ and $\hat{s}$ is an unbiased estimator of s.

Let $H\hat{s}(s,\mu)$ be the empirical histogram of $\hat{s}$ for fixed s and μ generated by simulations, then as can be seen from figures 5.10 and 5.11, the histogram $H\hat{s}(s,\mu)$ is centered at s and is practically independent of μ.

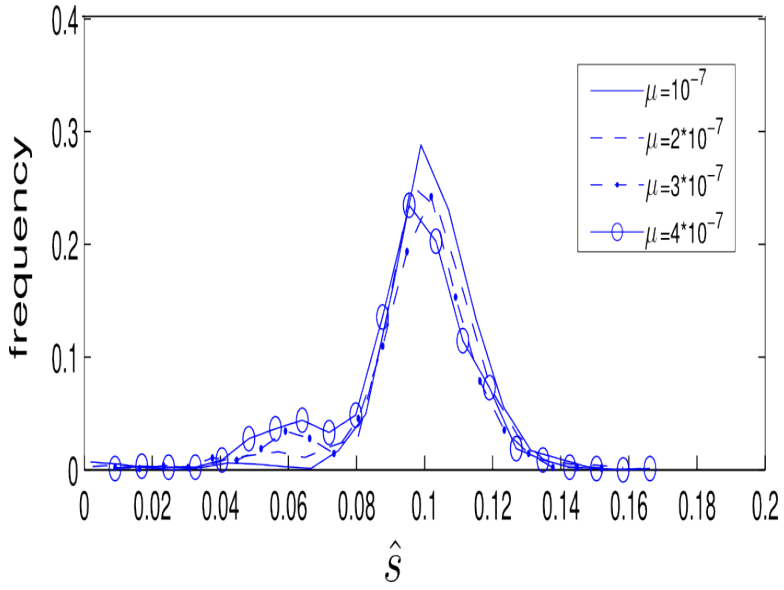


Figure 5.10: Empirical histograms of $\hat{s}$ for 4 values of μ are almost identical for a fixed $s = 0.1$, is independent of μ.

We plot the median of $\hat{s}$ as a function of s for two extreme values of μ, for $\mu = 2 \times 10^{-8}$ and 10^{-6} in figure 5.12, we can see that the new estimator of s based on marker frequencies estimated by complementary sub-sampling of size $N_{sub} = 300$ is not as accurately independent of μ as was the estimator $\hat{s}$ based on exact evaluations of marker frequencies, corresponding to ideal complementary sub-sampling size where N_{sub} would be equal to $N_0 = 50,000$ (figure 4.5). Thus the estimator $\hat{s}$ based on evaluations of marker frequencies by small size complementary sub-sampling is not

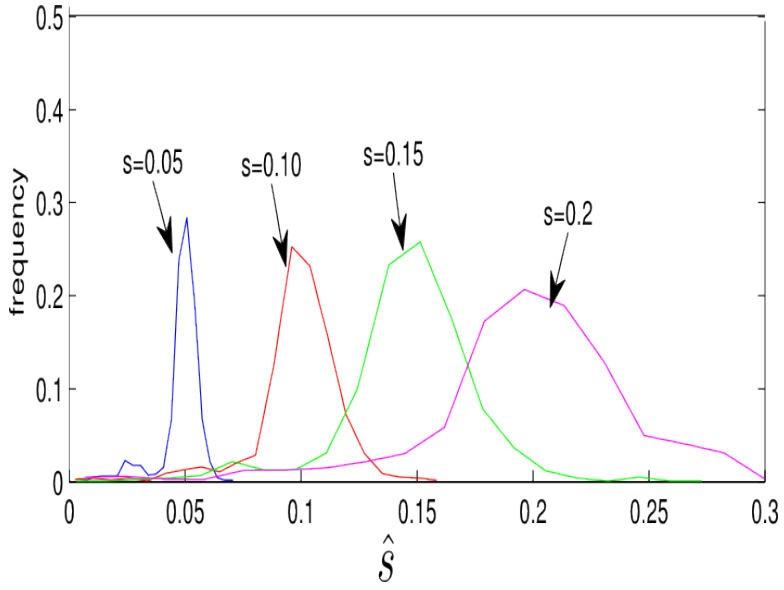


Figure 5.11: Empirical histograms of $\hat{s}$ are centered at s as displayed for 4 values of s and for a fixed $\mu = 2 \times 10^{-7}$.

completely unbiased.

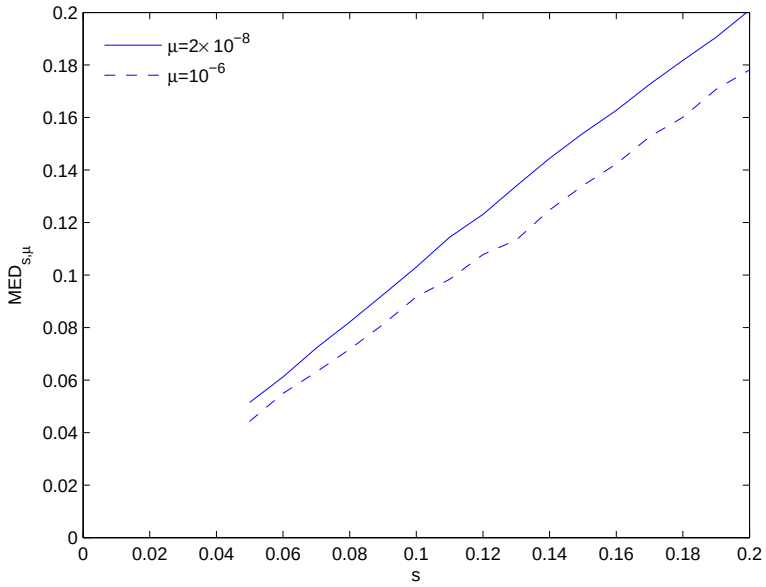


Figure 5.12: The empirical median of $\hat{s}$ as plotted for two extreme values of μ. The empirical median of $\hat{s}$ is approximately s but this reduces for large values of μ.

5.5.1 Accuracy of $\hat{s}$ and Comparison to the Experimental Data

As seen above, the estimator $\hat{s}$ of s is not completely an unbiased estimator of s with respect to μ, as was the case for the estimator of s when there is no complementary sub-sampling. We apply the above described algorithm to the experimental data, as these data gave numbers of red and white cells at the end of the day after the 2nd sub sampling had been performed. We compute an estimate, called the predicted estimates of s for each of the populations $Pop_1, \ldots, Pop_{\mathcal{N}}$. The direct experimental values $s_{dir}(j)$ for the selective advantage for each population Pop_j is obtained as before in Section 5.2. The predicted estimates $\hat{s}(j)$, using the simulation model, and the "ground truth" value $s_{dir}(j)$ of selective advantages are displayed below in the table 5.2.

The accuracy of the predicted estimates, using simulations, can be calculated by computing the square root of the mean square error $\sigma = \sqrt{E[\hat{s} - s]^2}$.

Next, to compare the accuracy of the Predicted $\hat{s}(i)$ and Observed $s_{dir}(i)$, we compare the two variances τ_i^2 corresponding to $s_{dir}(i)$ and σ_i^2 corresponding to $\hat{s}(i)$ for $i = 1, \ldots, 11$.

Let the variances τ_i^2 be known, then we want to compare to the variance σ_i^2. Let

$$V_1 = \tau_i^2 + \sigma_i^2$$

Table 5.2: Predicted estimates $\hat{s}$ and Observed direct experimental values s_{dir} of selective advantages for populations of experiments starting with ancestor cells.

Population	$\hat{s} \pm \sigma$	$s_{dir} \pm \tau$
Pop_1	0.14 ± 0.03	0.10 ± 0.004
Pop_2	0.09 ± 0.02	0.09 ± 0.005
Pop_4	0.13 ± 0.03	0.14 ± 0.005
Pop_5	0.08 ± 0.02	0.12 ± 0.006
Pop_6	0.13 ± 0.03	0.08 ± 0.004
Pop_7	0.11 ± 0.02	0.10 ± 0.008
Pop_8	0.08 ± 0.02	0.10 ± 0.009
Pop_9	0.11 ± 0.02	0.15 ± 0.005
Pop_{10}	0.13 ± 0.03	0.13 ± 0.004
Pop_{11}	0.11 ± 0.02	0.10 ± 0.004

be the combined variance and

$$V_2 = \frac{1}{10}\sum_{i=1}^{10}(\hat{s}(i) - s_{dir}(i))^2$$

be the mean of squared differences. Then to test the agreement of the direct and model based estimates, we calculated the statistic

$$X^2 = 10 * \frac{V_2}{V_1}.$$

Under our null hypothesis, each difference $s_{dir}(j) - \hat{s}(j)$ follows a distribution with mean zero and variance V_1, and X^2 is expected to follow a χ^2 distribution with 10 degrees of freedom. We reject the hypothesis that these two variances are almost the same, when

$$10 * \frac{V_2}{V_1} < \chi^2(\alpha/2, 10)$$

or

$$10 * \frac{V_2}{V_1} > \chi^2(1 - \alpha/2, 10).$$

Now, for the above case, we get that $10*\frac{V_2}{V_1} = 12.6582$. And if we check the hypothesis at 5% significance level for α, then we get $\chi^2(\alpha/2, 10) = 3.247$ and $\chi^2(1-\alpha/2, 10) = 20.483$. Here using the values $\chi^2(\alpha, 10)$ such that $\int_0^{\chi^2(\alpha,10)} f(x)dx = \alpha$. Thus we accept the hypothesis that the two variances are compatible. Hence, the predicted estimates using our estimation method, and the observed direct experimental values for s are still compatible.

5.5.2 Loss of Accuracy of $\hat{s}$ due to Complementary Sub-sampling

In this Section, we give a brief comparison between the estimators $\hat{s}$ when there is no complementary sub-sampling (Section 4.2.1), and when $\hat{s}$ is computed after the complementary sub-sampling is applied, as in Section 5.5. As seen before (in Section 4.2.1), when marker frequencies are ideally estimated with no error, our estimator $\hat{s}$ is computed by the following formula:

$$\hat{s}_{ideal} = \frac{1.26\ \hat{a} + 0.002}{5.29 - \hat{a}}. \tag{5.2}$$

When marker frequencies are estimated by daily complementary sub-sampling of size 300, our estimator $\hat{s}_{sub}$ of s is computed by the formula:

$$\hat{s}_{sub} = \frac{1.17\ \hat{a} + 0.005}{5.29 - \hat{a}}. \tag{5.3}$$

For each one of these two situations, fix $\mu = 5 \times 10^{-7}$. We compute the error of estimations Err for each s associated to μ, to see how far the true value of s is from the estimate $\hat{s}$. Call $\hat{s}_{ideal}$ the estimate of s based on ideal exact values of marker frequencies. Let $\hat{s}_{sub}$ denote the estimate of s based on marker frequencies

evaluated by complementary sub-sampling (extraction of approximately 300 cells from new culture wells and transfer of these 300 cells onto culture plates for visual cell counting). The mean estimation error $Err(s, \mu)$ for each s and when $\mu = 5 \times 10^{-7}$ is fixed, is defined by $Err(s, \mu) = mean(|\hat{s} - True\ s|)$ and calculated by empirical mean on simulated trajectories. Figure 5.13 displays the plot for these errors as calculated for the two estimators ($\hat{s}_{ideal}$ and $\hat{s}_{sub}$). We see that the accuracy for the estimator of s is reduced when the complementary sub-sampling is performed. While the $Err \leq 0.018$ for the estimator $\hat{s}_{ideal}$, we see that the Err for the estimator $\hat{s}_{sub}$ after the complementary sub-sampling increases to 0.035.

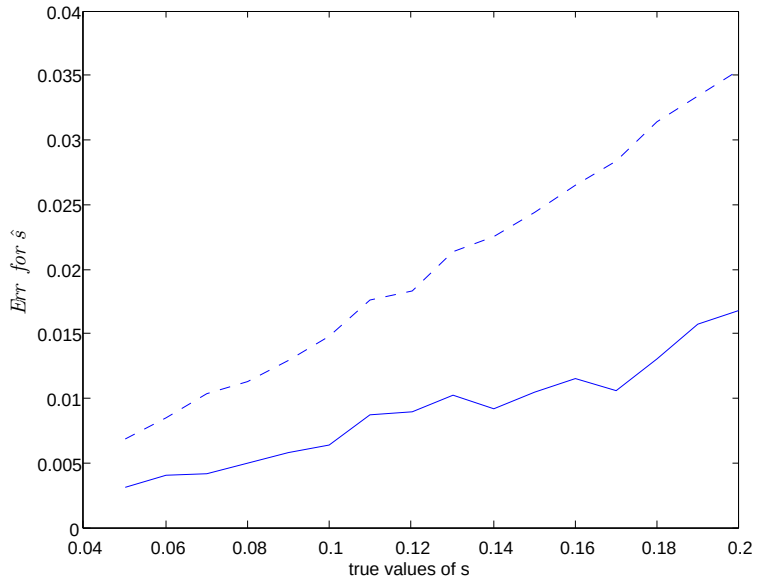


Figure 5.13: The plots for Err as computed for the estimator $\hat{s}_{ideal}$ when there is no complementary sub-sampling (in solid line) and Err as computed for the estimator $\hat{s}_{sub}$ based on frequencies estimated by complementary sub-sampling (in dashed line curve) when $\mu = 5 \times 10^{-7}$.

5.5.3 Accuracy of $\hat{s}$ for Different Sub-sampling Sizes

We display accuracy results for the estimator $\hat{s}$ as we increase the size N_{sub} of the daily complementary sub-samples. We compute the indicator $Err = mean(|\hat{s} - s|)$ for all s for a fix value of $\mu = 2 \times 10^{-7}$. We display the results for Err as we increase the size for $N_{sub} = 400,\ 1000,\ 5000,\ 10,000$, and $50,000$. This last value of N_{sub} corresponds to ideal exact evaluation of marker frequencies. Figure 5.14 displays these curves for Err for different N_{sub}. We see that the accuracy of estimator $\hat{s}$ increases as we increase the size N_{sub} of the daily complementary sub-samples.

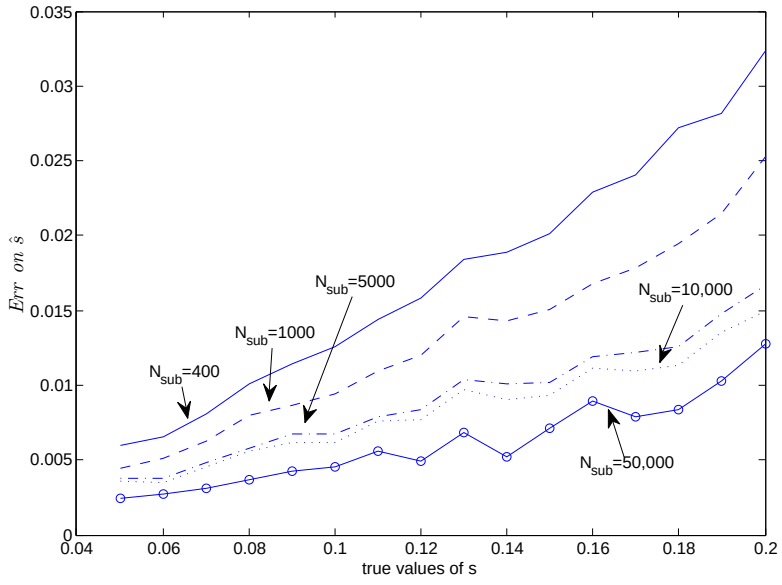


Figure 5.14: The curves for Err for the estimator $\hat{s}$ based on frequencies estimated by daily complementary sub-samples of size N_{sub} is displayed for all values of $s \in [0.05, 0.2]$ and for a fixed $\mu = 2 \times 10^{-7}$. The accuracy for the estimator increases as the size N_{sub} increases from 400, 1000, 5000, 10,000, to the maximal ideal size 50,000.

5.6 Accuracy for Estimator of Logarithmic Mutation Rate when Frequencies are Estimated by Complementary Sub-sampling

As mentioned above, the logarithmic mutation rate $\nu = log\mu$ cannot be measured directly in the TC nor in the HK experiments. But by simulations of the underlying process, we can compute accuracy for our estimator $\hat{\nu}$ of ν. Hence the computed accuracy of $\hat{\nu}$ cannot be evaluated by comparison with non existing experimental "ground truth values" of ν.

We link $\nu = \log \mu$ to the logarithm of the conditional probability P_{bot} of daily bottleneck crossing on day t given that there were no mutants at the beginning of that day. The range of values for $\nu = \log \mu$ is $(-18.42, 13.82)$. We estimate ν by the algorithm detailed above in Section 4.2.3. During the estimation of the logarithmic mutation rate, we treat the estimate $\hat{s}$ as the true value of the unknown parameter s. We first compute the intermediary estimator $\hat{\nu}_{int}$ and then form the final estimator $\hat{\nu}$ by using a re-centering technique similar to the technique presented in Section 4.2.8.

Here we define the re-centered estimator $\hat{\nu}$ of ν as the midpoint of the 90 % confidence interval for ν based on $\hat{\nu}_{int}$. We have indeed verified that in the context of frequency estimation by complementary sub-sampling, this simpler re-centering technique is more robust than the re-centering technique introduced above in Section 4.2.6.

Indeed in the context of complementary sub-sampling and for small number of wells $\mathcal{N} = 11$ the confidence intervals for ν based on $\hat{\nu}_{int}$ tend to be less precise as compared to the case when marker frequencies are assumed to be known exactly. The

adaptive re-centering weights outlined in 4.2.6, indeed tend to be less precise when frequencies are measured with strong accuracy, and thus often shift the final estimator $\hat{\nu}$ of ν too much to the right in the presence of complementary sub-sampling. For this adaptive weights technique, and in the presence of complementary sub-sampling, we display in figure 5.17 a typical example for the histogram of $\hat{\nu}$. Thus, when complementary sub-sampling is used, we definitely use the simpler re-centering by taking the midpoint of a confidence interval based on $\hat{\nu}_{int}$ which gives a robust accuracy for the final estimator $\hat{\nu}$.

The algorithmic computation of $\hat{\nu}_{int}$ only requires to know the observed values of the $\mathcal{N}$ times $T_\beta^1, \ldots, T_\beta^{\mathcal{N}}$, and the estimate $\hat{s}$ of s. Since the recorded frequency $p(t)$ is more corrupted by errors when frequencies estimates are based on complementary sub sampling, we consider a percentage β slightly higher than 55% to define the times T_β, namely $\beta = 60\%$, where $T_\beta = \inf\{t | p(t) > \beta\}$. To compute estimator accuracy for $\hat{\nu}$ given a fixed s and a fixed number $\mathcal{N}$ of replicate populations, we simulate 1000 values of $\hat{\nu}_{int}$ to generate an empirical histogram for the estimator $\hat{\nu}_{int}$. To compute confidence intervals $CI(\hat{\nu}_{int})$ for ν, we implement the algorithm outlined in chapter 4, and then one generates $\hat{\nu}$ as the midpoint of $CI(\hat{\nu}_{int})$.

The re-centered estimator $\hat{\nu}$ performs better than the preliminary estimator $\hat{\nu}_{int}$. Figures 5.15 and 5.16 display examples of histograms for the preliminary estimator $\hat{\nu}_{int}$ and for the final estimator $\hat{\nu}$. To evaluate the performance of the two estimators, as mentioned above, we compute the indicator Err. This is displayed in figures 5.18 for different values of s. Clearly, the accuracy is improved for the final estimator $\hat{\nu}$ as compared to the preliminary estimator $\hat{\nu}_{int}$, as can be seen in figure 4.15.

As could be expected, the accuracy for the estimator $\hat{\nu}$ of ν based on daily frequencies is decreased by the errors on frequencies due to the complementary sub-sampling and cell plating. When marker frequencies are measured "exactly", the mean absolute error of estimation Err for the estimator $\hat{\nu}$ of $\nu = \log \mu$ is less than 0.8 (figure 4.15) for $\mathcal{N} = 11$ and when $s = 0.10$ is fixed. In the presence of daily complementary sub-sampling of size $N_{sub} = 400$, the mean absolute error for estimator $\hat{\nu}$ increases to 1.5 in the same context ($\mathcal{N} = 11$ and $s = 0.10$).

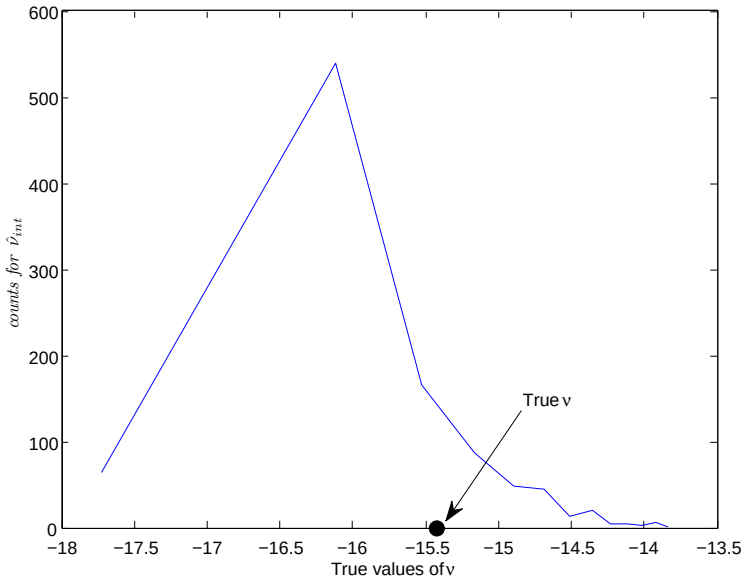


Figure 5.15: Displays the plot for the histogram of the preliminary estimator $\hat{\nu}_{int}$ of ν when $s = 0.10$, and $\mu = 2 \times 10^{-7}$ (or $\nu = -15.43$), and when $\mathcal{N} = 11$ is fixed. This tends to underestimate the true value of ν.

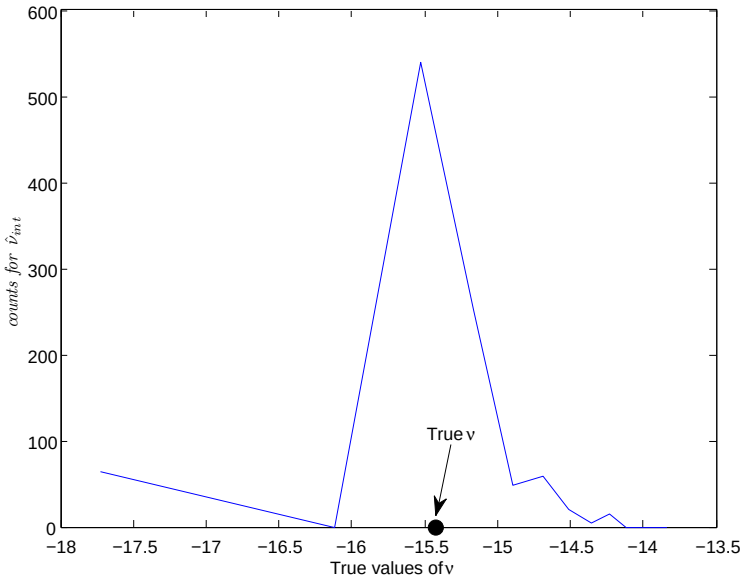


Figure 5.16: Displays the plot for the histogram of the final estimator $\hat{\nu}_{int}$ of ν when $s = 0.10$, and $\mu = 2 \times 10^{-7}$ (or $\nu = -15.43$), and when $\mathcal{N} = 11$ is fixed, which now is centered around the true value of ν.

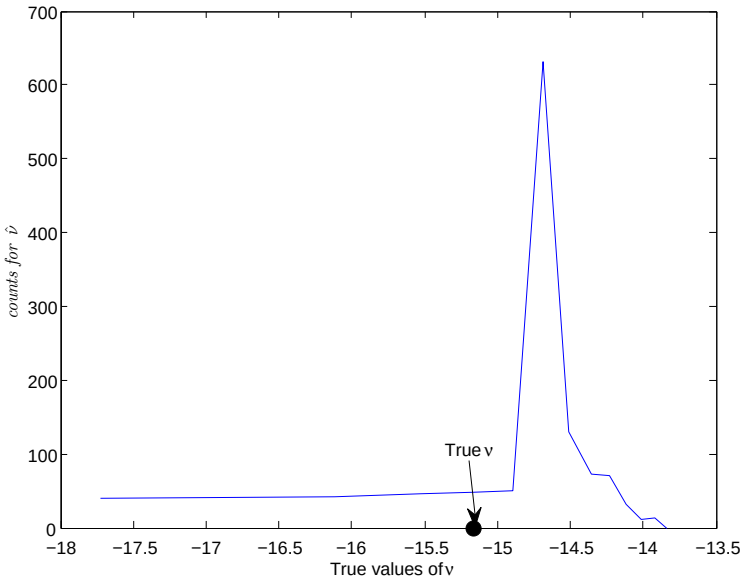
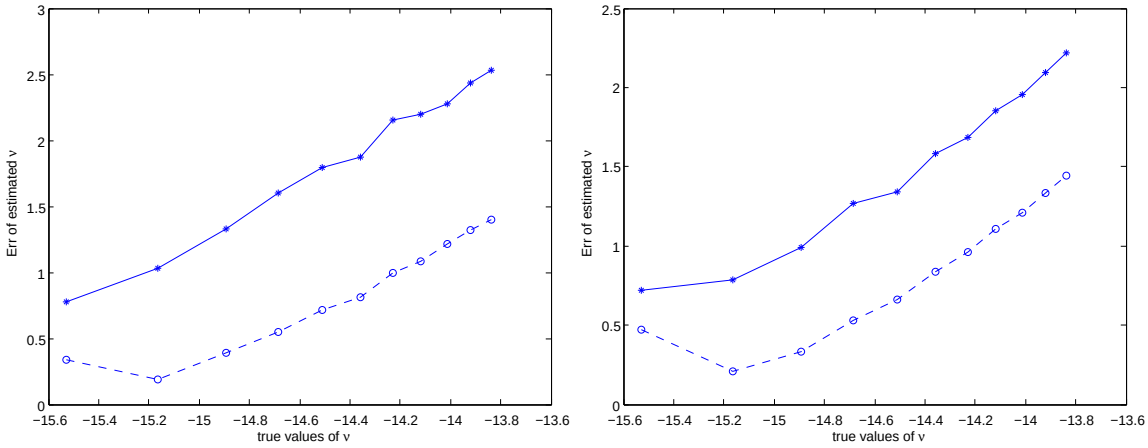


Figure 5.17: Displays the plot for the histogram of the final estimator (computed using the algorithm 4.2.6), when $s = 0.12$, $\mu = 2.6 \times 10^{-7}$ and when $\mathcal{N} = 11$, which does not work very well for estimation of ν after the complementary sub-sampling is performed.



(a) $s = 0.05$ (b) $s = 0.10$

Figure 5.18: The accuracy of the final re-centered estimator $\hat{\nu}$ is better than the accuracy for the intermediary estimator $\hat{\nu}_{int}$. This is displayed in figure (a) when $s = 0.05$ and in (b) when $s = 0.10$. The mean absolute errors of estimation are displayed by the solid line for the preliminary estimator $\hat{\nu}_{int}$ and by the dashed curve for the final re-centered estimator $\hat{\nu}$, as functions of the 13 values of $\nu = \log \mu$ present in our grid.

5.6.1 Accuracy of $\hat{\nu}$ for Different Sizes of the Complementary Sub-sampling

We next display the accuracy results for the estimator $\hat{\nu}$ as we increase the size N_{sub} of the daily complementary sub-samples. We display the results for the sizes $N_{sub} = 400, 1000, 5000, 10,000$ and $50,000$. The computation of the estimator $\hat{\nu}$ depends on the first time T_β when the frequency $p(t)$ reaches a percentage $\beta\%$. We see that when $\beta = 60\%$, the empirical histograms of T_β for different N_{sub} do not change much (see figure 5.19). The accuracy of our estimator $\hat{\nu}$ increases slightly as the size N_{sub} is increased. This is displayed in figure 5.20 where accuracy is quantified by mean squared errors.

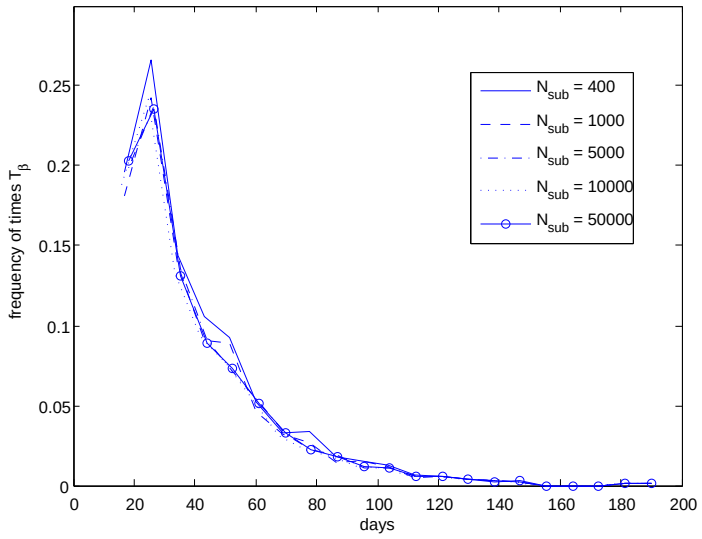


Figure 5.19: The plot for the empirical histograms of times T_β for different N_{sub}. These histograms do not change much when N_{sub} is modified when $\beta = 0.60$, $s = 0.10$ and $\mu = 2 \times 10^{-7}$.

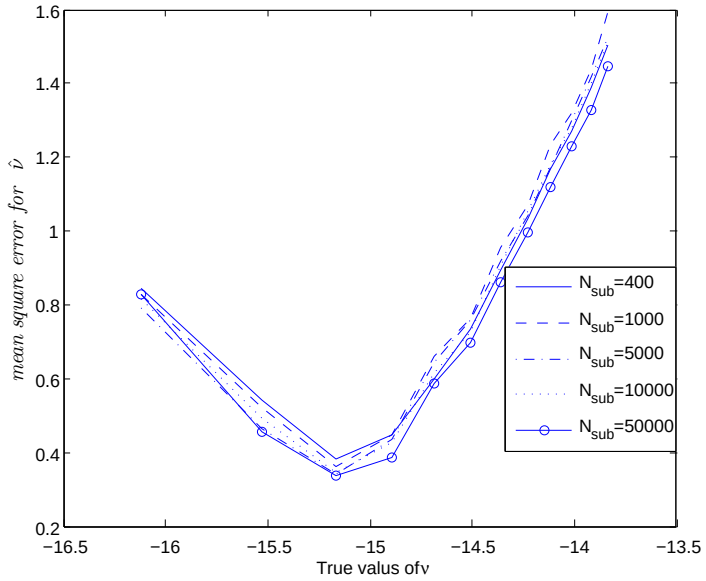


Figure 5.20: Plot of the mean squared estimation errors for $\hat{\nu}$ for different sizes N_{sub} of the daily complementary sub-samples , and $s = 0.10$. Even though the errors on $\hat{\nu}$ do not change much when $\beta = 0.60$, we can still see that mean squared errors decrease as the complementary sub-sampling size is increased from $N_{sub} = 400$ to $N_{sub} = 50000$. The circles denote the mean squared error of estimation for different true $\nu \in GRID$ values.

CHAPTER 6

Extension to Multiple Mutations

In this chapter, we extend the preceding study to models where we allow multiple types of mutations. Ancestor cells, as well as mutants, keep mutating further on. As before, we approximate the growth phase in 50 generations. The initial numbers for the two markers (red and white) are 2.5×10^4 respectively. Within the growth phase, mutations appear in both color markers, from ancestors as well as previous mutants, to reach a saturation growth capacity of $N_{sat} = 10^7$. The model, as before, consists of daily {growth + dilution} cycles.

6.1 Model

As before, we start with two initial sub-populations of the same initial genotype. Thus, we consider equal populations of red and white marker cells of identical "Ancestor Type". The total initial population of cells is 5×10^4. In the model, daily {growth + dilution} cycles are similar to those described above. The first daily phase is the "growth phase", while the second daily phase is the dilution process.

6.1.1 Growth Phase

The daily growth phase is a {growth + dilution} cycle. We divide a day into 50 time intervals, and hence each day involves 50 successive generations. Within the daily growth phase, the population well grows to a total maximum size of 10^7 cells until all nutrients have been consumed. The growth of red and of white cells within a day are modeled separately. Consider first the population of red cells. The growth phase for the population of white cells is modeled similarly. Consider a time interval $J(t)$ for $t = 1, \ldots, 50$. Let $m_r(t)$ be the number of different mutations or genotypes in the red cells at time $J(t)$. The population of red cells at the end of time interval $J(t)$ then consists of the initial ancestor genotype and the mutants that appeared until time interval $J(t)$, namely

$$(R_0(t), R_1(t), \ldots, R_{m_r(t)}(t))$$

where $R_i(t)$ for $i = 1, \ldots, m_r(t)$ denotes the size of genotype i, in the red population at the end of time interval $J(t)$, and $R_0(t)$ denotes the size of the ancestor genotype

in the red population at time interval $J(t)$.

Let s_i^R denote the selective advantage of cell type $R_i(t)$. The selective advantages of a cell type are computed with respect to the ancestor population, hence the selective advantage of the ancestor genotype is 0. Let $N_R(t) = \sum_{j=0}^{m_r(t)} R_j(t)$ be the size of the total number of red cells in the population at time interval $J(t)$. Occurrence of mutations during each growth time interval is random. The number of mutations at each time interval $J(t)$ within a day is distributed according to a Poisson distribution with parameter $\mu N_R(t)$, where the mutation rate μ is the input to the model.

Let $Nm_R(t)$ be the number of mutations occurring at time interval $J(t)$ in the red cells. These mutations occur in the different genotypes and the probability that a mutation occurs in the *ith* genotype is $R_i(t)/N_R(t)$. To assign these mutations, therefore, we do the following:

1. Create a vector $(p_0, \ldots, p_{m_r(t)})$ of length $m_r(t)+1$ so that $p_i = \frac{\sum_{j=0}^{i} R_j(t)}{N_R(t)} \quad \forall i = 0, \ldots, m_r(t)$. This creates a partition of the unit interval.

2. Let $RR = \{r_1, \ldots, r_{Nm_R(t)}\}$ where each $r_i \sim U(0,1)$ are random numbers sampled from the uniform distribution on the unit interval. Compute a vector $(Nm_R^0(t), \ldots, Nm_R^{m_r(t)}(t))$ where $Nm_R^i(t) = \#\{r \in RR : p_{i-1} \leq r \leq p_i\}$ for each $i = 0, \ldots, m_r(t)$.

Thus $Nm_R^i(t) = \#$ of mutations in genotype i in the red population at the end of time interval $J(t)$. To determine the selective advantages of these new emerging mutants, we explore different stochastic models. The selective advantages for the mutations depends on whether the mutation occurred from the ancestor genotype

or the mutation emerged from a previous mutant. To simulate the model described above, along with the mutation parameter, we need to specify

1. The density function or the histogram HA for random selective advantage assigned to mutants born from the ancestor genotype.

2. The density function or histograms HW for random selective advantage assigned to mutants born from a currently existing mutant.

The choice of HA and HW will define the stochastic model. If a new mutant is born from the ancestor, then its selective advantage is picked at random using the histogram HA. If a new mutant is born from an existing mutant population with selective advantage s_{ex}, then its selective advantage is sampled from the conditional density of HW, conditioned on the event $\{s > s_{ex}\}$.

Thus the set of genotypes present at time interval $J(t+1)$ is the union of the genotypes present at time t and the number of mutations that occurred at time $J(t)$. The new population is thus formed with selective advantages assigned as described above, to the new mutations that occur at each time interval within the growth cycle.

A similar model drives the population of white cells. Call $N(t) = N_R(t) + N_W(t)$ the total number of cells in the population at the end of time interval $J(t)$ (sum of red and white cells at the end of time interval $J(t)$). The daily {growth + dilution} cycle terminates when the total population reaches the maximum size of $N_{sat} = 10^7$ at the time τ when the nutrients in the wells are exhausted.

6.1.2 Dilution Phase

Suppose the growth phase terminates at time τ for $0 \leq \tau \leq 50$. At time τ, we have a population $Pop(\tau)$ of red and white cells. We extract a random sample of size $N_{sat}/200 = 5 \times 10^4$. The composition of the extraction of random sample has a multinomial distribution with parameters $S(\tau)$ = size of $Pop(\tau)$ and with probabilities $\frac{R_i(\tau)}{S(\tau)}$ and $\frac{W_j(\tau)}{S(\tau)}$ for $i = 0, \ldots, m_r(t)$ of red cells and $j = 0, \ldots, m_w(t)$ of white cells respectively. The randomly extracted population has size 5×10^4, and becomes the initial population for the next daily {growth + dilution} cycle.

This succession of {growth + dilution} daily cycles is performed until either the maximum number of days for which the experiments are recorded is reached, or $\frac{\sum_{j=0}^{m_r(t)} R_j(\tau)}{S(\tau)} \leq 0.01$ or $\frac{\sum_{j=0}^{m_w(t)} W_j(\tau)}{S(\tau)} \leq 0.01$, which corresponds to the event when the population of red or white cells captures the entire population and fixation of one color has occurred.

6.2 Different Models for Selective Advantages

Different models for assigning the selective advantages of mutants have been studied, for instance, by Hegreness et al. (2006) [23]. To assign the selective advantages to mutants, Hegreness et al. explore several distributions of beneficial mutations, including the exponential distribution, as suggested by Gillespie in (1991) [19]. The authors (Hegreness et al. (2006) [23]), show that very dissimilar underlying distributions: exponential, uniform, lognormal, and even an arbitrary distribution; all yield a similar distribution of successful mutations.

Based on this motivation, we explore different densities to draw the selective advantages of beneficial mutations, and to obtain the best fitting model to our experimental data (obtained from T. Cooper's laboratory). We explore

1. Model $E(\mu, \lambda)$ where the selective advantages are sampled from exponential distribution.

2. Model $EB(\mu, \lambda, a, b)$ where the selective advantages are sampled from exponential distribution on a bounded interval.

3. Model $EMP(Hist_{first}, Hist_{win})$ where the selective advantages are picked from histograms of estimated selective advantages of winner (using 6.5) from experimental data. This will be some arbitrary distribution for sampling the selective advantages of the beneficial mutations.

We discuss these models in detail below and in chapter 7.

6.2.1 Model "$E(\mu, \lambda)$": Exponential Densities for Selective Advantages

Here the histograms HA and HW both coincide with the exponential densities $E(\mu, \lambda)$ defined by $f(s) = \lambda \exp(-s\lambda) 1_{s>0}$ where $\lambda > 0$. The mean selective advantage is then given by $\frac{1}{\lambda}$.

We simulate the stochastic model defined above, where the random selective advantages assigned to emerging mutants have exponential densities with mean parameter $\frac{1}{\lambda}$. We systematically explore a whole range of potential values for λ, and

for the mutation rate μ.

6.2.2 Model "$EB(\mu, \lambda, a, b)$": Exponential Densities on a Bounded Interval

We have studied, by simulations, the stochastic evolution model described above, where the random selective advantage assigned to mutants have an exponential density restricted to a bounded interval. The histograms HA and HW above both coincide with an exponential density $EB(\mu, \lambda, a, b)$ on bounded interval, with density given by $f(x) = c(\lambda, a, b)e^{-x\lambda}1_{a<x<b}$ for $x \in I$, where I is the range of selective advantages. More explicitely, the density for such an exponential is redefined and given by

$$f_{a,b}(x, \lambda) = \frac{1}{\int_a^b \lambda e^{-\lambda\, x} dx} \lambda e^{-\lambda x} 1_{[a,b]}(x) \;=\; \frac{\lambda e^{-\lambda x}}{e^{-\lambda a} - e^{-\lambda b}} 1_{[a,b]}(x).$$

The mean $\bar{s}$ of this exponential is then given by:

$$\bar{s} \;=\; \frac{1}{\lambda} + \frac{ae^{-\lambda a} - be^{-\lambda b}}{e^{-\lambda a} - e^{-\lambda b}}.$$

6.2.3 Model "EMP": Based on Empirical Histograms

Our experimental data (Tim Cooper's laboratory) provides two empirical histograms HA and HW, which can be estimated from the experimental data. We use there data to compute a histogram $Hist_{first}$ for the estimated selective advantages of the mutants arising from the ancestor genotype. Fix $HA = Hist_{first}$ for each mutation scenario studied (which we will explain below). We also form histograms $Hist_{win}$ for

the estimated selective advantages of the winning mutations. Fix $HW = Hist_{win}$ for each mutation scenario studied. We will explain (in Section 6.5 below) the estimation of histograms $Hist_{first}$ and $Hist_{win}$.

6.3 Simulations of the Multiple Mutations Models

For stochastic models $E(\lambda)$ and $EB(\lambda)$ explained above, we explore a wide range of the parameters: λ (for the exponential distribution), and for the mutation rate μ. We fix $\mu \in \{10^{-7}, 2 \times 10^{-7}, \ldots, 10^{-6}\}$, and explore various values for λ for each model. For each pair of λ and μ (for the model used), we generate 1100 simulation trajectories. For each trajectory, some of the things that the simulations track include the times at which the mutations emerge, the selective advantage of each mutation, the color and genotype of the emerging mutation, the proportion of each genotype in the population.

6.4 Examples of Dynamic Evolution of Mutants

In this Section, we display some examples of the plots for the evolution of population. We display the trajectories for $g(t) = \log \frac{p(t)}{1-p(t)}$ for some cases. Then the evolution on mutations corresponding to such trajectories is displayed. These plots display dynamics for the emergence and fall of mutants frequency. We let simulations run until the maximum number of days considered for the growth of the population, and display the dynamics of mutants growth even after until fixation has occurred.

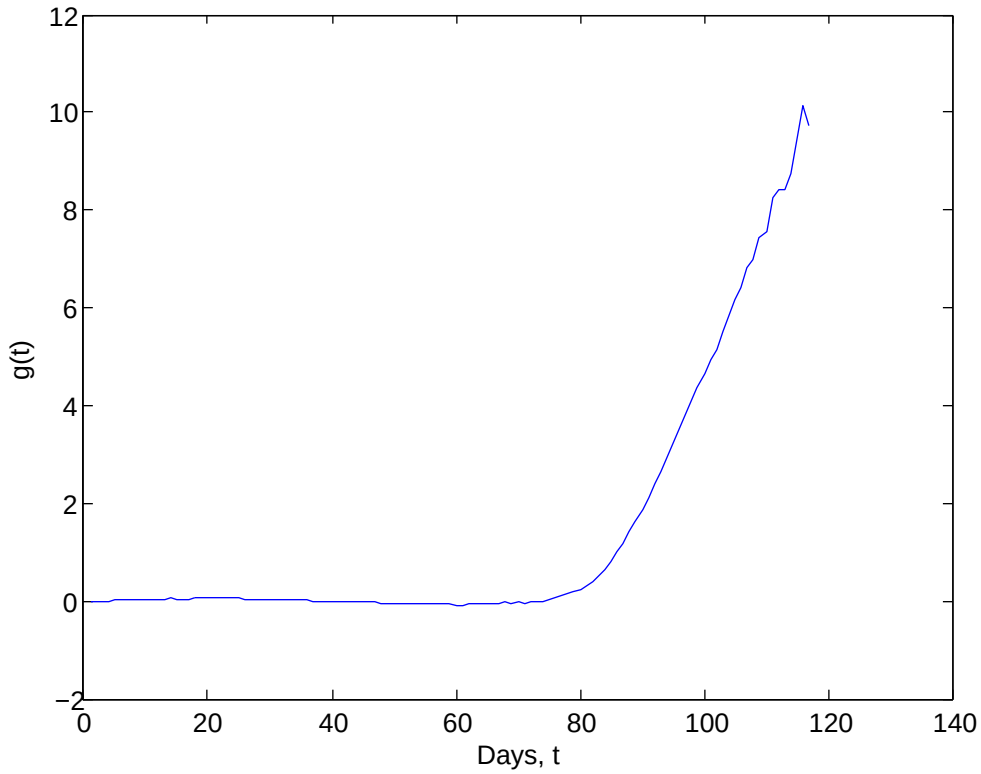


Figure 6.1: Example of the plot for the trajectory $g(t) = \log \frac{p(t)}{1-p(t)}$.

Figures 6.1 and 6.4 display examples of two such curves, along with the corresponding figures displaying the proportion of the mutants that emerged, displaying their dynamics, in figures 6.2 and 6.5 respectively. Only the mutants that emerged and stayed for a few days or have significant proportion in the population are displayed in these plots. The plots for the corresponding genealogy trees are also displayed (figures 6.3 and 6.6). These plots display the emergence of mutants, from ancestor or from previous mutants. These trees displays all the mutants that ever emerged in that particular population.

The first mutant emerges in the white population, from ancestor at time $T_1 = 55$

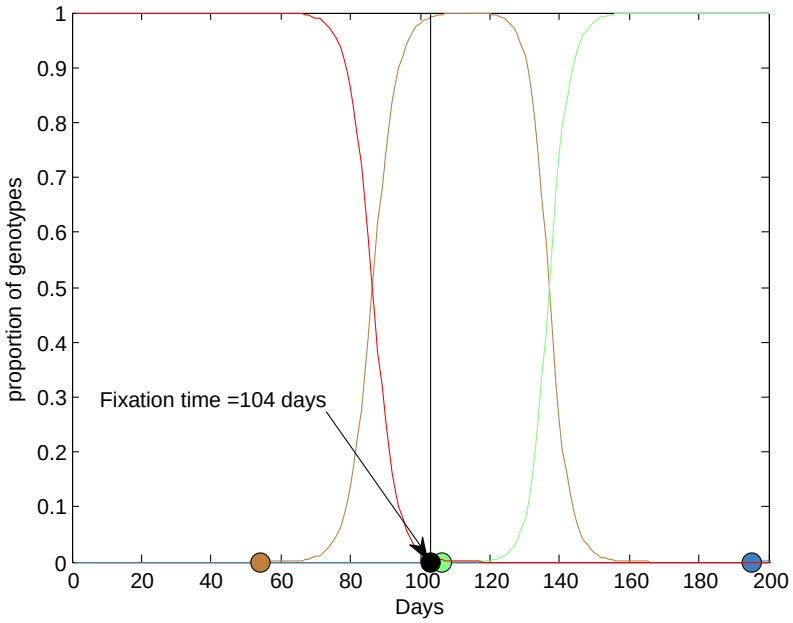


Figure 6.2: Dynamics for the evolution of the mutants that emerged in the population displayed in 6.1.

with selective advantage $s_1 = 0.05$. Thus proportion of ancestor genotype starts to decline and this being the only mutation, starts to increase. Right at the time of fixation, another mutation emerged (in white) from the previously existing mutant, with higher selective advantage $s_2 = 0.11$, causing mutant 1 to decrease in proportion. Third mutant emerges, in white (from mutant with selective advantage s_2) at day 195 (after fixation) with selective advantage $s_3 = 0.12$, but simulations terminate at day 200.

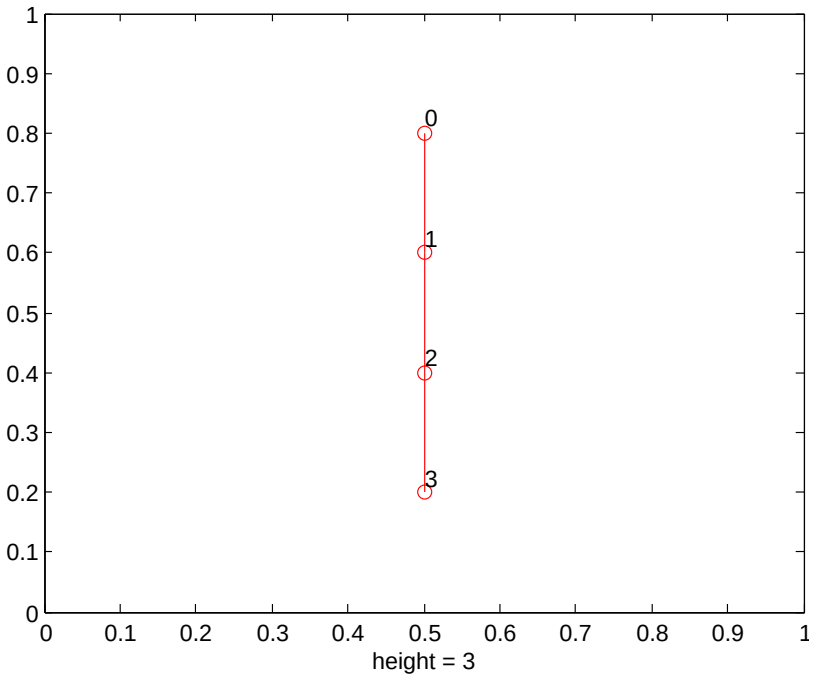


Figure 6.3: Genealogy trees displaying the order of emergence of mutants. The height of the tree is 3. Note that y-axis is not needed here.

Mutant 1 emerges in white with a selective advantage of $s_1 = 0.05$ and drags the trajectory $g(t)$ to fixation. Two other mutations, as seen in figure 6.2 also emerge in white, but after the day at which fixation. Mutant 2 with higher selective advantage 0.11 emerges from 1, and mutant 3 with selective advantage 0.12 emerges from 2.

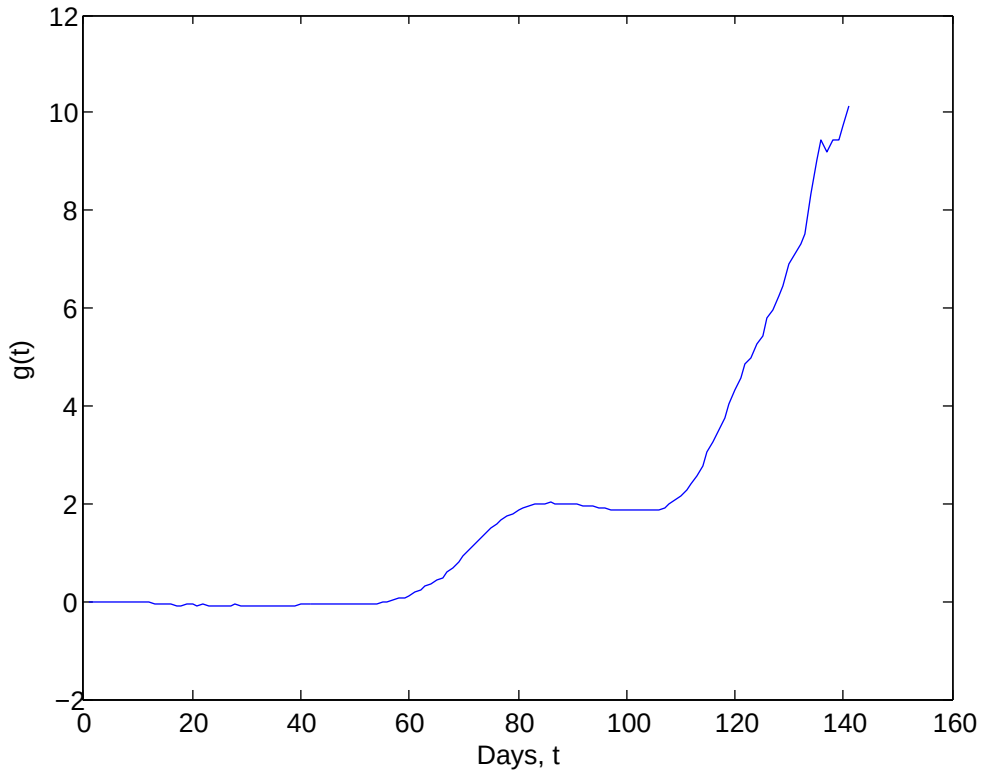


Figure 6.4: Example of the plot for the trajectory of $g(t) = \log \frac{p(t)}{1-p(t)}$.

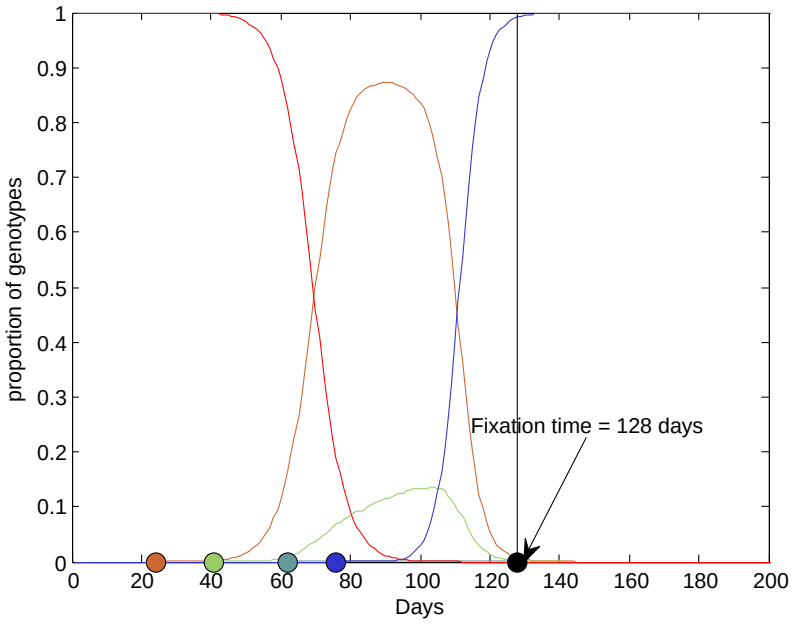


Figure 6.5: Dynamics for the evolution of the mutants that emerged in the population trajectory displayed in 6.4.

Mutants emerge in both red and white cells and compete with each other. First mutation emerges in white at day 22 with selective advantage $s_1 = 0.04$ and starts to increase until emergence of another mutant in red population at day 40, with equal selective advantage $s_2 = 0.04$. The increase of this mutation is slow as its selective advantage is equal to the existing mutation. Yet another mutant with almost same advantage $s_3 = 0.04$ appears in white at day 61 but does not increase in proportion. This mutation is then followed by another mutant emerging at day 78 in white population with higher selective advantage $s_4 = 0.1$. This mutation drags the population to fixation. We see that it takes longer for the population to go to fixation due to emergence of more weaker mutants, and due to mutants emerging in both red and white cells.

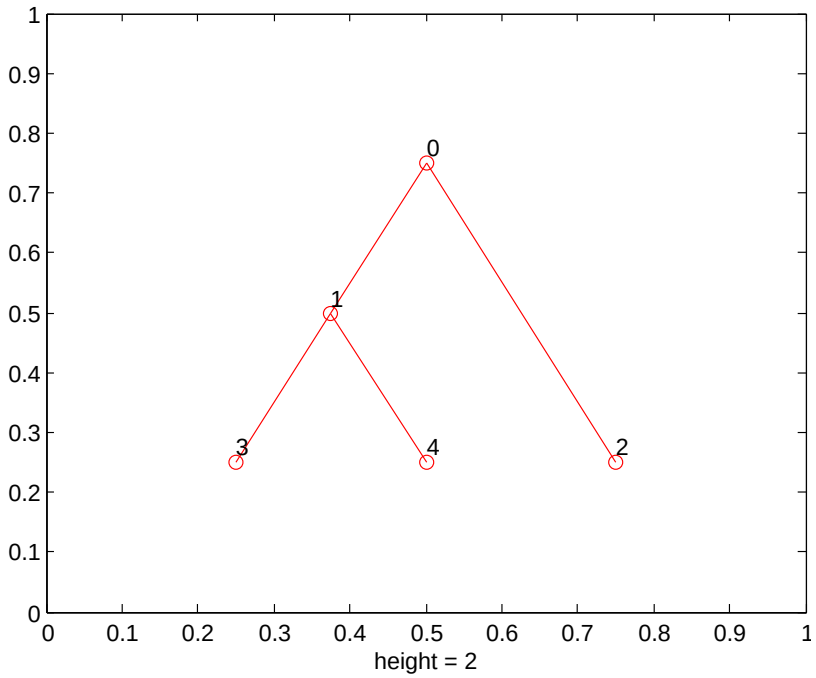


Figure 6.6: Plots for the genealogy trees displaying the order of emergence of mutants. The height of this tree is 2. Note that the y-axis is not needed for this tree plot.

Initially mutants with almost equal selective advantage appear in both red and white cells, until mutant 4 with higher selective advantage 0.10 emerges in white from an existing mutant 1 in white and drags the population to fixation.

6.5 Estimation of Histograms $Hist_{first}$ and $Hist_{win}$

In our research group, a non-linear least squares modeling algorithm has been developed and implemented by Wei Zhang to fit the observed experimental data with the iterative deterministic systems for each mutation scenario. This is mentioned here for completeness. Let $r_0(t), r_1(t)$ and $r_2(t)$ be the day t frequencies of red ancestors, mutants occurring from ancestors and mutants arising from the previous mutants, respectively. Let $w_0(t)$, $w_1(t)$ and $w_2(t)$ be the day t frequencies of the white ancestors, mutations occurring from ancestors, and mutants occurring from the previous mutants, respectively. Let s_1 and s_2 be the respective selective advantages of the mutants arising from the ancestors, and for mutants arising from the previous mutants, respectively. And let μ_1 and μ_2 be the corresponding mutation rates for these mutants. The multiplicative growth factor per time interval for the ancestors is $F = D^{1/m}$ where $m = 50$ is the number of time intervals during a day growth, and D is the dilution factor. The multiplicative growth factor per time interval for the mutants born from the ancestor genotype is then given by $M = F^{1+s_1}$, and the multiplicative growth factor for the mutants born from existing previous mutants is given by $G = F^{1+s_2}$.

For a given observed trajectory, for each possible scenario for occurrence of mutations, one fits the experimental data by the iterative deterministic system, and apply non linear least squares fitting to obtain estimates for the unknown selective advantages s_1 and s_2.

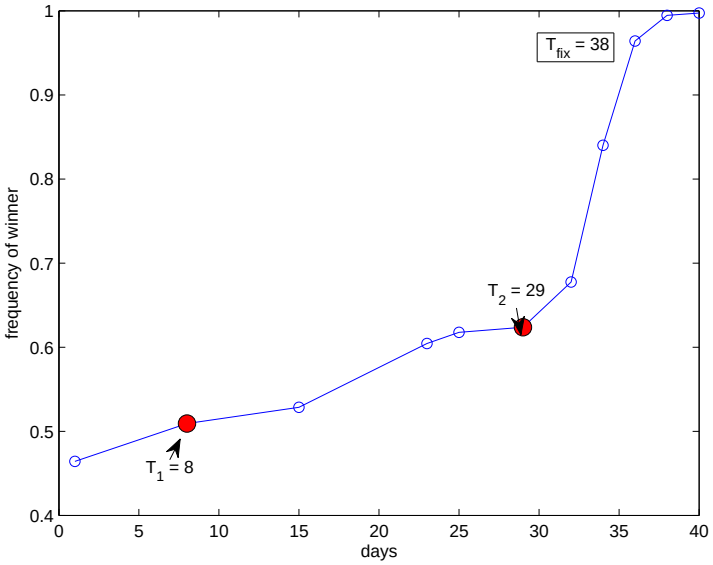


Figure 6.7: An example of the frequency of winner plot for a population generated by "ancestor".

From figure (6.7), we see that this population trajectory has two mutants emerging at times T_1 and T_2. The plot displayed is for the frequency of the white marker cells. We form the following two scenarios:

1. Scenario A: Emergence of first mutant M_1 at time $T_1 = 8$. This has the selective advantage as estimated to be $s_1 = 0.01$. The second mutant M_2 emerges at time $T_2 = 29$ with a selective advantage of $s_2 = 0.15$. The second mutant also arises from the progenitor or the ancestor. This mutant with a higher selective advantage drags the population to fixation, which occurs at time $T_{fix} = 38$.

2. Scenario B: The first mutant M_1 emerges at time $T_1 = 8$ with a selective advantage $s_1 = 0.01$ as estimated using the method described above of fitting by non-linear least squares. Second mutant further emerges from the previous M_1 mutant at time $T_2 = 29$ with a selective advantage of $s_2 = 0.15$ as computed

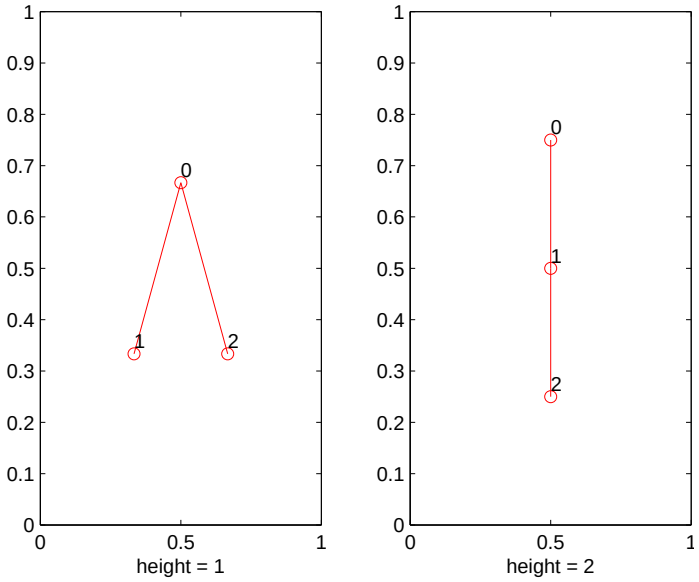


Figure 6.8: Possible genealogy trees of mutants.

with respect to the progenitor. This drags the population to fixation at time $T_{fix} = 38$.

The possible genealogy trees of the two mutants that emerged in this particular population are displayed in figure 6.8. For trajectories fitted with multiple scenarios, an F test was applied to evaluate the quality of fit.

Given a trajectory, the red and white marker frequencies $r(t)$ and $w(t)$ can be extracted and are thus known. The iterative deterministic system describing the average growth dynamics is specified by the following three equations

$$w_0(t+1) = w_0(t)\, F^{n_t}/D \tag{6.1}$$

$$w_1(t+1) = w_1(t)\, M^{n_t}/D \tag{6.2}$$

$$r_0(t+1) = r_0(t)\, F^{n_t}/D. \tag{6.3}$$

The constraints are given by

$$w(t+1)\ D = w_0(t)\ F^{n_t} + w_1(t)\ M^{n_t} + [w(t) - w_0(t) - w_1(t)]\ G^{n_t} \quad (6.4)$$

$$r(t+1)\ D = r_0(t)\ F^{n_t} + [r(t) - r_0(t)]\ M^{n_t}. \quad (6.5)$$

The integer solution for n_t can be solved from the following equations

$$(w_0(t) + r_0(t))\ F^{n_t} + (w_1(t) + r_1(t))\ M^{n_t} + w_2(t)\ G^{n_t} = D.$$

Note that this modeling algorithm does not require knowing the mutation rate μ, and estimates the mutation scenario parameters $s_1, s_2, \ldots$ separately for each observed process trajectory. The distinction between which mutation scenario best fits a particular population or in which population some mutation scenarios are equivalent is determined by applying the F test as explained above. The details for this can be seen in the joint paper to appear. For each experiment trajectory (obtained from Tim Cooper's laboratory), we obtain estimates for the selective advantages of the emerging mutations, and thus generate histograms H_{first} and H_{win}.

CHAPTER 7

Fitting Multiple Mutation Models to Experimental Data

In this chapter, we present a set of statistical tests to quantify the quality of fit between multiple mutation models and experimental data. The goal is to develop a method for evaluating the fitting of multiple mutation models to experimental data. We will present this generic method and apply it to 6 different sets of data with $\mathcal{N}$ populations each. The difference between these 6 experiments is the genotype of the initial $\mathcal{N}$ identical populations. The list of initial genotypes for the 6 sets of data is below:

1. Experiment 1: consists of the initial genotype with no mutation present at the

beginning: The Ancestor genotype.

2. Experiment 2: consists of the initial genotype with the 1st kind of mutation, called the Ribose.

3. Experiment 3: consists of the initial genotype with 1st and 2nd kind of mutations, namely, Ribose and TopA. Thus experiment 3 has initial genotype with 2 mutations, called RiboseTopA.

4. Experiment 4: consists of the initial genotype with 1st, 2nd and the 3rd kind of mutations, namely, Ribose, TopA, and spoT. Thus experiment 4 has initial genotype with 3 mutations, called RiboseTopAspoT.

5. Experiment 5: consists of the initial genotype with 1st, 2nd, 3rd, and 4th kind of mutations, namely, Ribose, TopA, spoT, and glmus. Thus experiment 5 has initial genotype with 4 mutations, called RiboseTopAspoTglmus.

6. Lastly, Experiment 6: consists of the initial genotype with 1st, 2nd, 3rd, 4th, and 5th kind of mutations, namely, Ribose, TopA, spoT, glmus, and pykF. Thus experiment 6 has initial genotype with 5 mutations, called RiboseTopAspotglmuspykF.

For each of the $\mathcal{N}$ populations, the fixation time T_{fix} is defined as $T_{fix} := \inf\{t|p(t) > 0.99\}$, i.e, the first time the frequency of winner exceeds 99%. For each of the $\mathcal{N}$ observed population trajectories, as defined in chapter 6(6.5), in our research group, we compute the selective advantages of the mutants present in that trajectory.

7.1 Strategy for Fitting Models to Data

7.1.1 Simulations of Multiple Mutation Models

Given a set of $\mathcal{N}$ experimental trajectories recording the evolution of $\mathcal{N}$ identical initial populations, we will use intensive simulations of the multiple mutation models introduced in chapter 6. We allow ancestors, as well as mutants, to mutate further. We start with two sub-populations of the same initial genotype. The size of the initial population is $N_0 = 5 \times 10^4$. The model undergoes daily {growth + dilution} cycles until a population reaches a maximum size of $N_{sat} = 10^7$, when the nutrients in the wells are exhausted. The selective advantages of the mutants depends on whether the mutation occurred from the ancestor genotype or the previous mutant, as explained in chapter 6. Call $\mathcal{M}$, the multiple mutation model and Θ the set of parameters determining the model $\mathcal{M}$. The model parameters are the mutation rate μ, and the parameters of the density function defining the selection of selective advantages. We explore different models based on selection of selective advantages as introduced in chapter 6, namely,

1. Model "$E(\mu, \lambda)$": Exponential densities for selective advantages. Let this represent a model $\mathcal{M}(\Theta)$ with parameters $\Theta = (\mu,\ \lambda)$, with $\lambda > 0$ being the parameter for the exponential.

2. Model "$EB(\mu, \lambda, a, b)$": Exponential densities on a bounded interval. Let this represent a model $\mathcal{M}(\Theta)$ with parameters $\Theta = (\mu,\ \lambda,\ a,\ b)$, where $a,\ b$ are the end-points for the bounded interval considered, and λ is the parameter for the

exponential.

3. Model EMP: based on empirical densities, and hence let this represent a model $\mathcal{M}(\Theta)$ with parameters $\Theta = (\mu,\ Hist_{first},\ Hist_{win})$.

For each set of parameters, and hence each parameterized model $\mathcal{M}(\Theta)$, we generate 1000 trajectories and store the simulation data base for further exploration steps.

7.1.2 $Test_{fix}$: Comparison of Fixation Times

For each simulated parameterized model $\mathcal{M}(\Theta)$, we obtain by simulations, a set of 1000 trajectories. From this simulation data base, we can compute the empirical histograms for the fixation times. Call these fixation times to be T_{sim}. Further, as stated above, we also compute the fixation times for each of the $\mathcal{N}$ observed populations. The experimental data give us a list of $\mathcal{N}$ experimental values for the fixation times (call these $T^1_{obs}, \ldots, T^{\mathcal{N}}_{obs}$). Let $T_{min,\,obs} = \min\{T^1_{obs}, \ldots, T^{\mathcal{N}}_{obs}\}$, be the minimum of these $\mathcal{N}$ observed fixation times, and let $T_{max,\,obs} = \max\{T^1_{obs}, \ldots, T^{\mathcal{N}}_{obs}\}$, be the maximum of these $\mathcal{N}$ observed fixation times. We calculate the probability $p(\Theta) = Pty(T_{min,\,obs} < T_{sim} < T_{max,\,obs})$, that the fixation times generated by the model $\mathcal{M}(\Theta)$ falls between $T_{min,\,obs}$ and $T_{max,\,obs}$. Since we are observing $\mathcal{N}$ independent populations, the probability that $T_{min,\,obs} < T_{sim} < T_{max,\,obs}$ for all $\mathcal{N}$ trajectories is $p(\Theta)^{\mathcal{N}}$. Hence at significance level $\alpha = 5\%$, we will accept the hypothesis $(\Theta = \Theta_0)$ when $p(\Theta_0)^{\mathcal{N}} \geq \alpha = 5\%$.
We explore this for a wide range of the values for the parameters Θ.

7.1.3 $Test_{swin}$: Comparison of Histograms for the Selective Advantages of the Winner

We want to compare the histograms for s_{win}, the selective advantages of the winner, from simulated and observed $\mathcal{N}$ trajectories. For each of the above simulated parameterized model $\mathcal{M}(\Theta)$, we obtain by simulations, a set of 1000 trajectories. From this simulation data base, we compute the empirical histograms $hist_{sim}(\Theta)$, for the selective advantages of the winner. Using the 1000 simulated s values, we create histogram with 41 bins that we smooth in the manner described below. To smooth this histogram, for each s value, we take 4 bin centers to the left of this s value, and 4 to the right, and define the density at s to be the average of the density over these 8 bins. For the first 4 bins, and for the last 4 bins, we set the density equal to the density on the left (or right) end-points. Further, as stated in the beginning of this chapter and in chapter 6, we estimate the selective advantages of the winner, (using non-linear least squares fitting, Section 6.5) for each of the $\mathcal{N}$ observed populations; call this set $Z = (Z_1, \ldots, Z_{\mathcal{N}})$. Let $f_\Theta(Z_i)$ represent the density of Z_i derived from the estimated densities. Then the likelihood function $L_\Theta(Z) = \prod_{i=1}^{\mathcal{N}} f_\Theta(Z_i)$ is given as the product of the $\mathcal{N}$ densities as observed in the histogram $hist_{sim}(\Theta)$. Thus the log-likelihood of observing Z is

$$LL_\Theta(Z) = \log(L_\Theta(Z)) = \sum_{i=1}^{\mathcal{N}} \log(f_\Theta(Z_i)).$$

The average log-likelihood $V_\Theta(Z)$ is then defined by

$$V_\Theta(Z) = \frac{1}{\mathcal{N}} LL_\Theta(Z).$$

This represents the average log-likelihood value of the true experimental estimates of s_{win}. We now create another histogram of the average log-likelihood values by extracting a random set of $\mathcal{N}$ selective advantages from the simulated histogram $hist_{sim}(\Theta)$ of s_{win} and computing the corresponding average log-likelihood $V_\Theta(Z)$ as above. That is, we form $V_\Theta(X^1), \ldots, V_\Theta(X^{500})$ where $X^i = (X_1^i, \ldots, X_{\mathcal{N}}^i)$ for $i = 1, \ldots, 500$, is the set of $\mathcal{N}$ values sampled from the histogram $hist_{sim}(\Theta)$ and then compute $V_\Theta(X^i) = \frac{1}{\mathcal{N}} LL_\Theta(X^i))$. Note that the density for each X_j^i for $i = 1, \ldots, 500$ and for $j = 1, \ldots, \mathcal{N}$ is computed with respect to the simulated histogram of s_{win}. This generates an empirical histogram for the average log-likelihood values. Let $VV_{sim}(\Theta) = (V_\Theta(X^1), \ldots, V_\Theta(X^{500}))$.

For the estimates of s_{win} extracted from the experimental data, we have $\mathcal{N}$ trajectories and thus $\mathcal{N}$ estimates for s_{win}, and a corresponding true value $V_\Theta(Z)$. From the above empirical histogram of the average log-likelihood values $VV_{sim}(\Theta)$, we form the two quantiles $Q_- = 5\%$ quantile and $Q_+ = 95\%$ quantile. If our true average log-likelihood $V_\Theta(Z)$ value lies inside the quantile range obtained from simulations, i.e, $Q_- \leq V_\Theta(Z) \leq Q_+$, then we accept the hypothesis Θ.

The best quality of fit to experimental data is based on our probability $p(\Theta)$ and the result based on our second test $Test_{swin}$.

We also computed the empirical distribution of the number of mutants in the winning branch, and the simulated densities of the number N_{win} of mutants in the winning branch. When this number $N_{win} = 2$, this means that a first mutant is born from the progenitor and the second mutant born from the previous mutant. This probability in simulations is low. Table 7.1 displays the results for this probability

$Prob(N_{win} = 2)$ in simulations when model$E(\mu, \lambda)$ is used to assign selective advantages, for different pairs of μ and the mean selective advantage $\bar{s}(\Theta) = \frac{1}{\lambda}$. We have much often winning mutants born directly from the progenitor.

Table 7.1: Displaying probability $Prob(N_{win} = 2)$ for different pairs of μ and mean selective advantage $\bar{s}(\Theta)$ for model $E(\mu, \lambda)$.

$\mu \backslash \bar{s}(\Theta)$	0.03	0.05	0.11
4×10^{-7}	0.10	0.12	0.11
8×10^{-7}	0.23	0.19	0.17
10^{-6}	0.23	0.22	0.18

7.2 Study of Experiment 1

In this Section, we study experiment 1 where starting initial genotype= Ancestor genotype, where no mutation has yet occurred. We explore quality of fit to data for the models described above in chapter 6.

By convention, the winning color marker is called white. The fixation times as computed from the trajectories of the frequency of white markers for the ancestor experimental populations is displayed in table 7.2. We also display the estimates for selective advantage s_{win} of the winner as obtained by the non-linear least squares algorithm explained in 6.5.

7.2.1 Exponential Density: Model $E(\mu, \lambda)$

This multiple mutation model $\mathcal{M}(\Theta)$ is parameterized with model parameters $\Theta = (\mu, \lambda)$ with $\lambda > 0$. We explore systematically a wide range of values for μ and λ, and

Table 7.2: The estimates of s_{win} and fixation times T_{fix} as obtained for each population of Experiment 1.

Population	s_{win}	T_{fix}
Pop_1	0.13	32
Pop_2	0.16	40
Pop_3	0.13	54
Pop_4	0.13	32
Pop_5	0.10	40
Pop_6	0.15	45
Pop_7	0.10	38
Pop_8	0.07	36
Pop_9	0.13	25
Pop_{10}	0.15	38
Pop_{11}	0.09	52

hence for the mean selection advantage $\bar{s}(\Theta) = \frac{1}{\lambda}$. The selective advantages of the mutants are randomly defined using an exponential density function as explained previously (chapter 6). The selective advantages of the mutations occurring from the progenitor are picked at random from the exponential density with parameter λ, and the selective advantages of the mutations occurring from the previous mutants is picked from the exponential distribution conditional on the fact that the selective advantage of this second mutant is greater than the selective advantage of the first mutant (as in chapter 6). The mean selective advantage $\bar{s}(\Theta)$ for this model is given by $\bar{s}(\Theta) = \frac{1}{\lambda}$. The simulated selective advantages of winners for this model are much higher than the observed range for s_{win}. We obtain the best μ and best $\bar{s}(\Theta)$ as follows:

1. Compute the pairs (μ, λ) which maximize the probability $p(\Theta)$ for $Test_{fix}$.

2. From these pairs, then compute the pair for best (μ, λ) by using $Test_{swin}$, as the

pair which maximizes the true $V_\Theta(Z)$, average log-likelihood for the observed s_{win}.

In this way, we compute the best μ and $\bar{s}(\Theta)$. From the table below (7.3) we see that, even for the best μ and $\bar{s}(\Theta)$ (in bold), one cannot reach a significant level of even 1% for the p-value of the fit between observed and models based fixation times density function. Also, we display the histograms obtained from simulation data corresponding to the best pair of μ and $\bar{s}(\Theta)$ (in bold). We display these histograms for s_{win}, and we see that the simulated selective advantages obtained are much higher than the observed s_{win} from the experiments. Figure 7.1 displays this histogram. The range for the experimentally observed selective advantages of winner, s_{win} is indicated in this figure by the green bar.

Thus the model $E(\mu, \lambda)$ with parameters $\Theta = (\mu, \lambda)$ does not fit very well, the experimental data for the ancestor genotype.

Table 7.3: Quality of fit for model $E(\mu, \lambda)$ with parameters $\Theta = (\mu, \lambda)$. The estimate of μ, and the mean selective advantage is in bold.

λ	Mean selective advantage	Best μ	$p(\Theta)^{\mathcal{N}}$
27	0.037	8×10^{-7}	7×10^{-5}
22	0.045	10^{-6}	2×10^{-3}
20	0.05	10^{-6}	4×10^{-3}
18	0.056	10^{-6}	5×10^{-3}
12.5	**0.08**	$\mathbf{8 \times 10^{-7}}$	$\mathbf{5 \times 10^{-3}}$
9	0.11	5×10^{-7}	10^{-3}
4.5	0.2	2×10^{-7}	4×10^{-4}

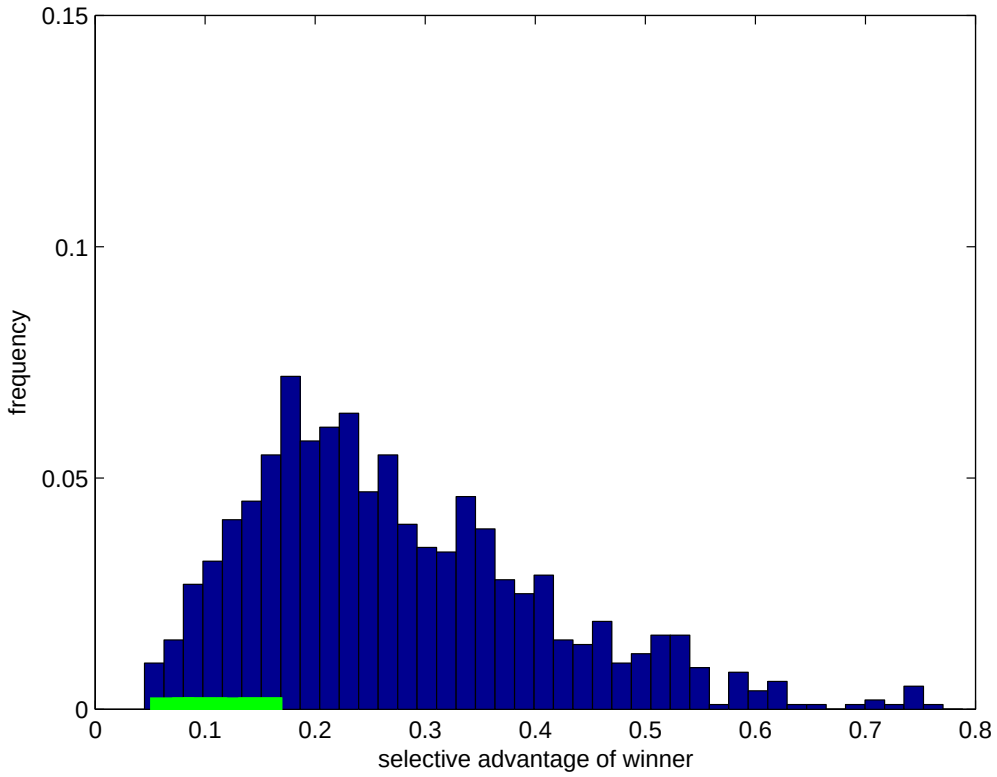


Figure 7.1: The empirical histogram of the selective advantage of the winner for the best exponential density $Exp(12.5)$ with $\mu = 8 \times 10^{-7}$. The green bar indicates the range $[0.05, 0.17]$ covered by the $\mathcal{N} = 11$ experimentally observed selective advantages of winner.

7.2.2 Model $EB(\mu, \lambda, a, b)$: Exponential Density on Bounded Interval

We see from the above multiple mutation model $\mathcal{M}(\Theta)$ that the free exponential density tends to over estimate the selective advantages of the winner. For this reason, we use the exponential density but restricted on a bounded interval. We form the model $\mathcal{M}(\Theta)$ with parameters $\Theta = (\mu,\ \lambda,\ a,\ b)$, where μ denotes the mutation rate and $\lambda > 0$ denotes the parameter for the exponential on a bounded interval defined by $[a, b]$. Here we pick the selective advantages of the mutants arising from the progenitor from $EB(\mu, \lambda, a, b)$, the exponential density restricted on a bounded interval $[a, b]$ while the selective advantage of the mutants arising from the previous mutants are picked from the exponential conditioned on the interval $[s_{existing}, b]$ where $s_{existing}$ is the selective advantage of the existing mutant from which the new mutation occurred. The determination of the bounded interval is made such that the observed s_{winner} all lie in this interval. We explore systematically a wide range of values for μ and λ. We also explore different intervals, namely $[a, b]$ for $a = \{0.01, 0.02, 0.03, 0.04\}$ and for $b = \{0.14, 0.15, 0.16\}$, but we did not observe any significant of changing the interval for the bounded exponential. The best μ and the best parameters Θ are obtained as explained above in 7.2.1. We obtain the best $\mu = 10 \times 10^{-7}$ for the exponential on a bounded interval $[0.01, 0.16]$ with parameter $\lambda = 1.5$. The best mean selective advantage is $\bar{s}(\Theta) = 0.08$. The probability $p(\Theta) = P(T_{min,\,obs} \leq T_{sim} \leq T_{max,\,obs}) = 0.78$, which when raised to the power $\mathcal{N} = 11$, gives a p-value of 6×10^{-2}. For this data, we get $T_{min,obs} = 23$ and $T_{max,obs} = 56$. We have a small probability $Pty(T_{sim} > 56) = 0.15$ and $Pty(T_{sim} < 23) = 0.08$. Figures 7.2 and 7.3

displays the empirical histogram for selective advantage s_{win} for the winner, and for the fixation times for the best model.

The empirical histogram of the average log-likelihood values is displayed in figure 7.4 below. The red dots indicate the quantiles Q_- and Q_+ of the average log-likelihood as obtained from the observed data. The green dot indicates the average log-likelihood $V_\Theta(Z)$ as obtained from the estimates s_{win} of selective advantage of winner from the experimental trajectories.

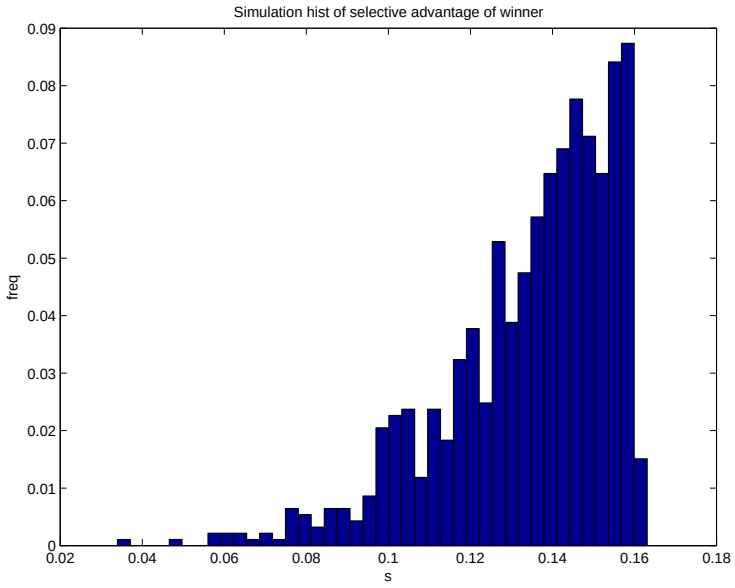


Figure 7.2: The empirical histogram for the selective advantage, s_{win} of the winner. Here the model $\mathcal{M}(\Theta)$ is based on $\Theta = (10 \times 10^{-7}, 1.5, 0.01, 0.16)$ for exponential density on a bounded interval. Here $\bar{s}(\Theta) = 0.08$.

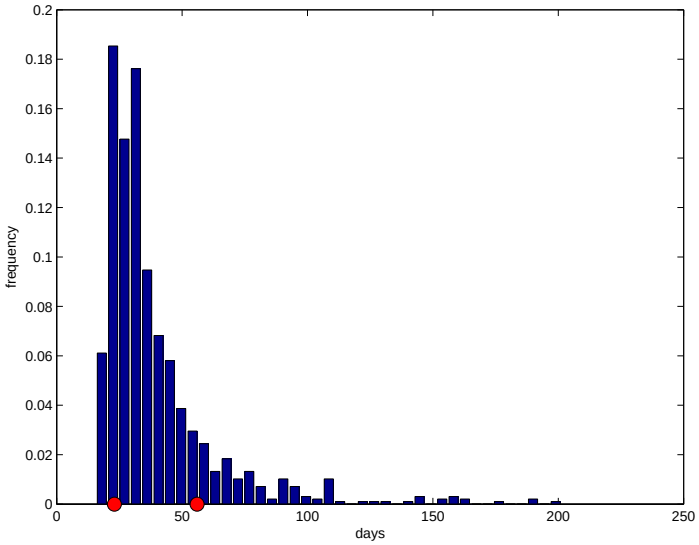


Figure 7.3: The empirical histogram for the fixation times. Here the model $\mathcal{M}(\Theta)$ is based on $\Theta = (10 \times 10^{-7}, 1.5, 0.01, 0.16)$ for exponential density on a bounded interval. Here $\bar{s}(\Theta) = 0.08$, and $p(\Theta)^{\mathcal{N}} = 6\%$. The red dots indicate the min and max of fixation times observed from the experimental data directly.

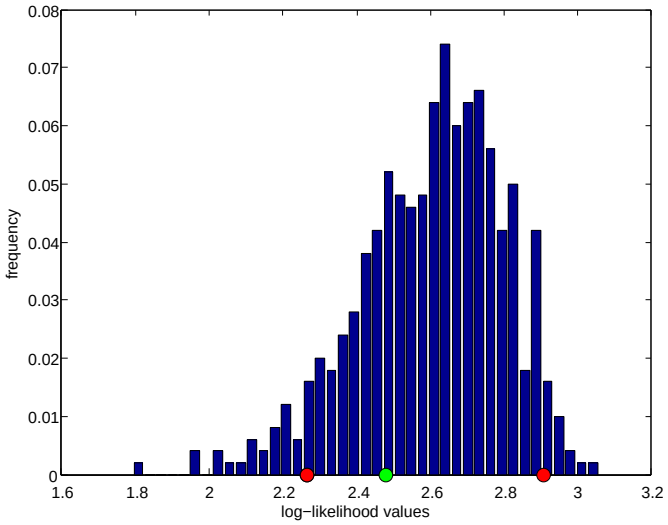


Figure 7.4: The empirical histogram average log likelihood values. Here the model $\mathcal{M}(\Theta)$ is based on $\Theta = (10 \times 10^{-7}, 1.5, 0.01, 0.16)$ for exponential density on a bounded interval. Here $\bar{s}(\Theta) = 0.08$. The true average log-likelihood pf observed s_{win} lies inside the quantiles Q_- and Q_+ of the simulated average log likelihood computed from s_{win} values.

7.2.3 Multiple Mutation Models Based on Empirical Densities

For the ancestor data set, model $\mathcal{M}(\Theta)$ where $\Theta = (\mu, Hist_{first}, Hist_{win})$ represents the multiple mutation model based on empirical densities. We pick random selective advantages for the mutants by using the empirical histograms that we obtain by studying the mutation scenarios for the ancestor data for the numbers of red and white marker type cells, as explained in chapter 6 (Section 6.5). We estimate histograms H_{first} consisting of the selective advantages of the mutants arising from the progenitor. By studying the mutation scenarios, we see that we have 6 trajectories out of $\mathcal{N} = 11$ with 2 possible mutation scenarios. This generates a total of $2^6 = 64$ different choices for the histogram H_{first}. Lets call these as 64 different hypothesis histograms that we have to study, since we will have these many different choices for the input histograms HA. We use Kolmogorov-Smirnov test at 5% significance level to test if hypothesis HA_i and HA_j for $i,\ j = 1, \ldots, 64$ are the same hypothesis. Thus, we obtain 14 different hypothesis in which we can combine the 64 hypothesis histograms. Lets call these 14 different groups for HA. For each group, we have a list of hypothesis histograms that are similar. We consider the average histogram of these similar hypothesis histograms to represent our histogram HA corresponding to that group, thus generating 14 different hypothesis histograms for HA. However, the empirical histogram for the selective advantage s_{win} of the winner is fixed for all the hypotheses. We study these different 14 groups and apply the two statistical tests ($Test_{fix}$ and $Test_{swin}$) explained above, to obtain the best group hypothesis histogram that fits the observed data. We compute the corresponding $p-$ values. For

different groups, these results are displayed in the table 7.4 below. The group maximizing these p-values obtained from the computations of $Test_{fix}$ and $Test_{swin}$, is called the best group hypothesis. Group 4 hypothesis turns out to be the best group with the probabilities $p(\Theta) = P(23 \leq T_{sim} \leq 56) = 0.77$ and $P(T_{sim} > 56) = 0.16$ and $P(T_{sim} < 23) = 0.07$. This corresponds to a $p-$ value of 6%.

Table 7.4: This table displays the results for the best μ that we obtain for the different hypothesis. The best μ and group is in red.

Group	best μ	$p(\Theta)$[11]
1	7×10^{-7}	3×10^{-2}
2	9×10^{-7}	3×10^{-2}
3	10^{-6}	3×10^{-2}
4	$\mathbf{10^{-6}}$	$\mathbf{6 \times 10^{-2}}$
5	10^{-6}	4×10^{-2}
6	8×10^{-7}	2×10^{-2}
7	9×10^{-7}	2×10^{-2}
8	9×10^{-7}	4×10^{-2}
9	9×10^{-7}	2×10^{-2}
10	10^{-6}	3×10^{-2}
11	9×10^{-7}	2×10^{-2}
12	10^{-6}	3×10^{-2}
13	10^{-6}	2×10^{-2}
14	9×10^{-7}	1×10^{-2}

The best model in quality of fit to the ancestor experimental data would be when we use exponential density on a bounded interval. We display the three best density plots below in figure 7.5.

Table 7.5: Displays different multiple mutation models and their quality of fit. The model $EB(\mu, \lambda, a, b)$ shows the best quality fit, with the parameters below.

Model	$p(\Theta)^{11}$	μ	Mean s_{win}
Best Model $E(\mu, \lambda)$	5×10^{-3}	8×10^{-7}	0.08
Best $EB(\mu, \lambda, 0.01, 0.16)$	6×10^{-2}	10×10^{-7}	0.08
Best Empirical Histogram	6×10^{-2}	10×10^{-7}	0.12

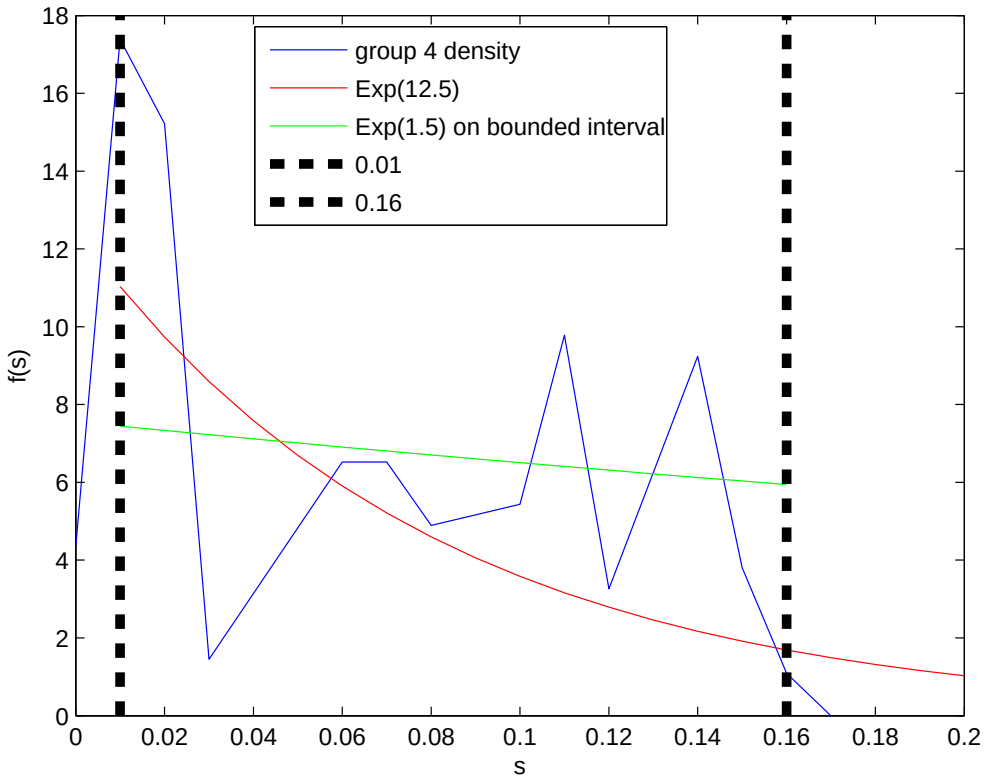


Figure 7.5: The density plot for the hypothesis group 4, the density plot for the best exponential $Exp(12.5)$, with mean 0.08 and the density plot for the best exponential with parameter $\lambda = 1.5$ on bounded interval $[0.01, 0.16]$, and hence with mean selective advantage 0.082.

7.3 Study of Experiment 2

In this Section, we study the 2nd set of experiments, that were started with the initial genotype as the genotype with one mutation, Ribose. From the above exploration, we see that using the model $E(\mu, \lambda)$ with exponential density for assigning the selective advantages of mutants , over estimates the selective advantages of the winner. Using the model $Emp(\mu, Hist_{first}, Hist_{win})$ with empirical density hypothesis or using the model $EB(\mu, \lambda, a, b)$ with exponential density on a bounded interval are kind of equivalent to study the experimental data. For this data set and others, we restrict our study to using the model $EB(\mu, \lambda, a, b)$ with exponential on a bounded interval. We choose the interval such that the observed estimates all lie inside that interval. We explore systematically a wide range of values for μ and λ, and hence for the mean selection advantage $\bar{s}(\Theta) = \frac{1}{\lambda}$. We display the result for the best pair that we obtain (as mentioned in above, 7.2.1) on maximizing the probability for $Test_{fix}$, and for the comparison of the selective advantages, s_{win} of the winner, using the test $Test_{swin}$. For this data set, the observed estimates of selective advantages of winner and the times it takes for the genotype to reach fixation are as follows displayed in Table 7.6. The estimates for the selective advantage are estimated with respect to their progenitor.

The best model estimates that we obtain for the mutation rate with this experimental data are for the exponential density $EB(\mu, \lambda, a, b)$ with parameter $\lambda = 13$ on the bounded interval $[0.01, 0.19]$. The estimate for μ is 9×10^{-7} and the mean selective advantage is $\bar{s}(\Theta) = 0.07$. We get a very good fit to the model, which

Table 7.6: The estimates of selective advantages and the fixation times as obtained for each population

Population	s_{win}	T_{fix}
Pop_1	0.11	34
Pop_2	0.03	80
Pop_3	0.17	32
Pop_4	0.16	32
Pop_5	0.15	34
Pop_6	0.10	25
Pop_7	0.17	38
Pop_8	0.04	42
Pop_9	0.17	90
Pop_{10}	0.12	36
Pop_{11}	0.08	52

can be seen from the figures as well as by the probabilities obtained by applying the above tests. The $T_{min,\,obs} = 23$ and $T_{max,\,obs} = 92$. Thus the probability $Pty(23 \leq T_{sim} \leq 92) = 0.89$, which when raised to the power $\mathcal{N} = 11$ gives a p-value 0.24. The empirical histogram of the times to fixation is displayed in figure 7.6 below. The empirical histogram obtained for the selective advantages of the winner for this best model is also displayed (figure 7.7). The comparison of the s_{win} using the average log-likelihoods test $Test_{swin}$ can be seen from the histogram displayed in figure 7.8. The plot for the exponential density with mean selective advantage 0.07 is also displayed.

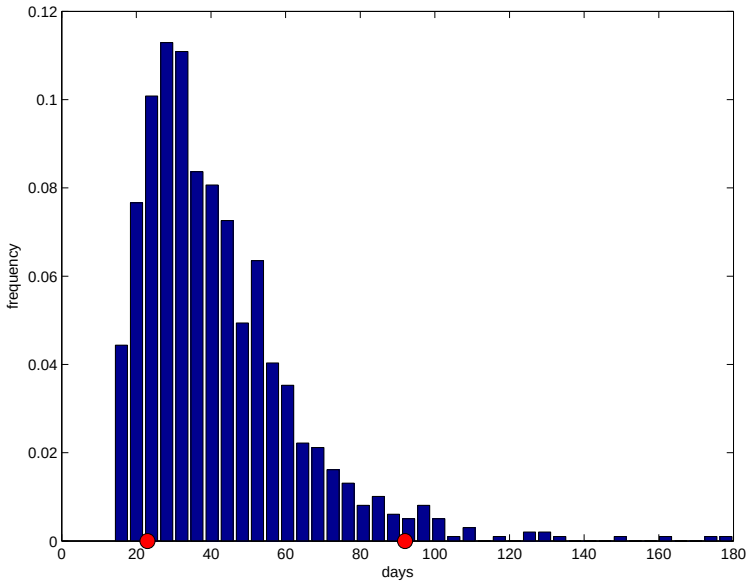


Figure 7.6: The empirical histogram of the fixation times for the best model $EB(\mu, \lambda, a, b)$, exponential density on a bounded interval $[0.01, 0.19]$ with $\mu = 9 \times 10^{-7}$, and mean selective advantage 0.07. The red dots indicate the min and max of fixation times observed from the experimental data directly. We obtain $p(\Theta)^{11} = 0.24$.

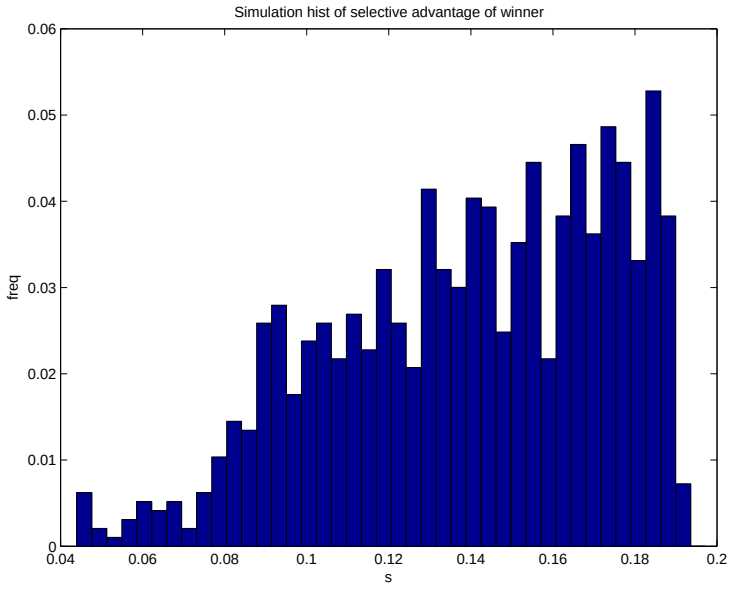


Figure 7.7: The empirical histogram of the selective advantage of the winner for the best model $EB(\mu, \lambda, a, b)$, exponential density on a bounded interval $[0.01, 0.19]$ with $\mu = 9 \times 10^{-7}$, and mean selective advantage 0.07.

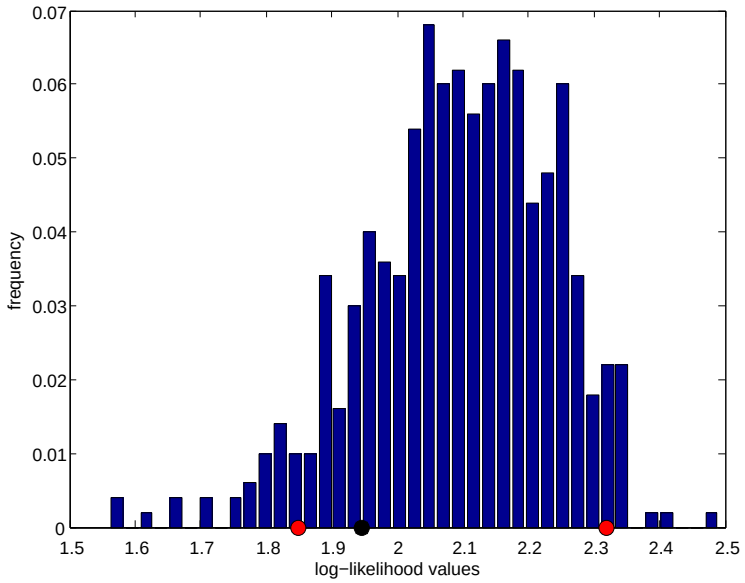


Figure 7.8: The empirical histogram average log-likelihood values for the best model $EB(\mu, \lambda, a, b)$ with $\mu = 9 \times 10^{-7}$, and mean selective advantage 0.07, on bounded interval $[0.01, 0.19]$. The true average log-likelihood pf observed s_{win} lies inside the quantiles Q_- and Q_+ of the simulated average log likelihood computed from s_{win} values.

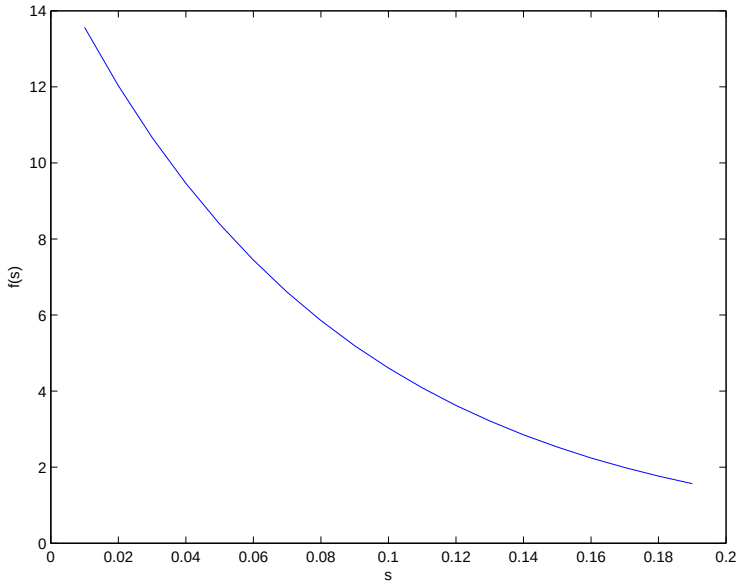


Figure 7.9: The density of selective advantages is plotted for best exponential density with parameter $\lambda = 13$ on bounded interval $[0.01, 0.19]$ and hence with mean selective advantage 0.07.

7.4 Study of Experiment 3

In this Section, we study the 3rd set of experiments, that were started with the initial genotype consisting of the genotype with 2 types of mutations, namely, Ribose and TopA. We compute the estimates of the selective advantages, s_{win} of winner. This is computed as explained before (6.5) by performing the non-linear least squares fitting to the populations. We again use the model $EB(\mu, \lambda, a, b)$ with exponential density on a bounded interval $[a, b]$ to assign the selective advantages of the mutants. We apply the tests ($Test_{fix}$ and $Test_{swin}$) explained above to systematically explored values of μ and λ and hence for mean $\bar{s}(\Theta)$. We display the result for the best pair of $\bar{s}(\Theta)$ and μ that we obtain by maximizing the probability $p(\Theta) = P(T_{min,\,obs} \leq T_{sim} \leq T_{max,\,obs})$ using $Test_{fix}$. We also perform the log-likelihood comparison, $Test_{swin}$ for the selective advantages of the winner. The observed estimates that we obtain for the selective advantages of winner for this data set are displayed in the table 7.7 below. The fixation times observed from each population trajectory for frequency of winner are also displayed. One population out of the $\mathcal{N} = 11$ populations, in this case, did not go to fixation, so we ignore that population from this study.

The best model estimate for the mutation rate was obtained for model $EB(\mu, \lambda, a, b)$ exponential density with parameter $\lambda = 30$ on a bounded interval $[0.02, 0.16]$. The estimate of μ is 10^{-6} and mean selective advantage is 0.05. The minimum $T_{min,\,obs} = 30$ and the maximum $T_{max,\,obs} = 73$. The probability for the range test $p(\Theta) = P(30 \leq$

Table 7.7: The estimates of selective advantages and the fixation times as obtained for each population

Population	s_{win}	T_{fix}
Pop_1	0.07	50
Pop_2	0.13	40
Pop_3	0.03	71
Pop_4	0.15	32
Pop_5	0.09	38
Pop_6	0.08	48
Pop_7	0.12	58
Pop_8	0.11	32
Pop_9	0.10	38
Pop_{10}	0.04	50

$T_{sim} \leq 73) = 0.72$ which when raised to 11 power gives a p-value of 0.03. The empirical histogram for the times to fixation is displayed in figure 7.10. The empirical histograms obtained from simulations for the selective advantage of the winner is also displayed (figure 7.11). The histogram for the average log-likelihood values is also displayed (figure 7.12). The best exponential density on a bounded interval with mean $\bar{s}(\Theta) = 0.05$ is also displayed (figure 7.13).

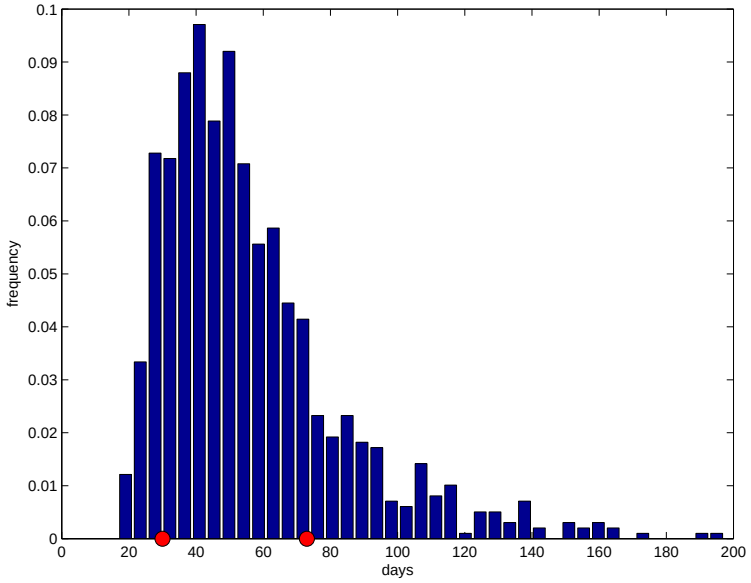


Figure 7.10: The empirical histogram of the fixation times for the best $EB(\mu, \lambda, a, b)$ exponential density on a bounded interval [0.02,0.16] with $\mu = 10 \times 10^{-7}$ and $\bar{s}(\Theta) = 0.05$. The red dots indicate the min and max of fixation times observed from the experimental data directly.

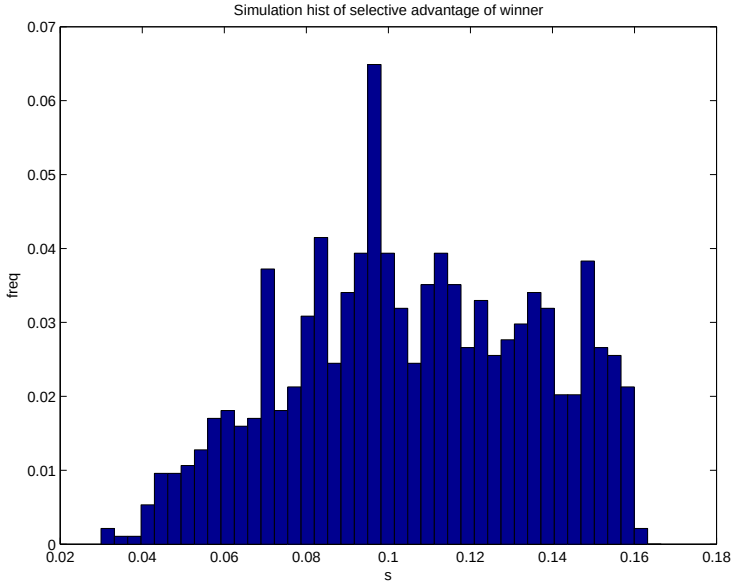


Figure 7.11: The empirical histogram of the selective advantage of the winner for the best $EB(\mu, \lambda, a, b)$ exponential density on a bounded interval [0.02,0.16] with $\mu = 10 \times 10^{-7}$ and $\bar{s}(\Theta) = 0.05$.

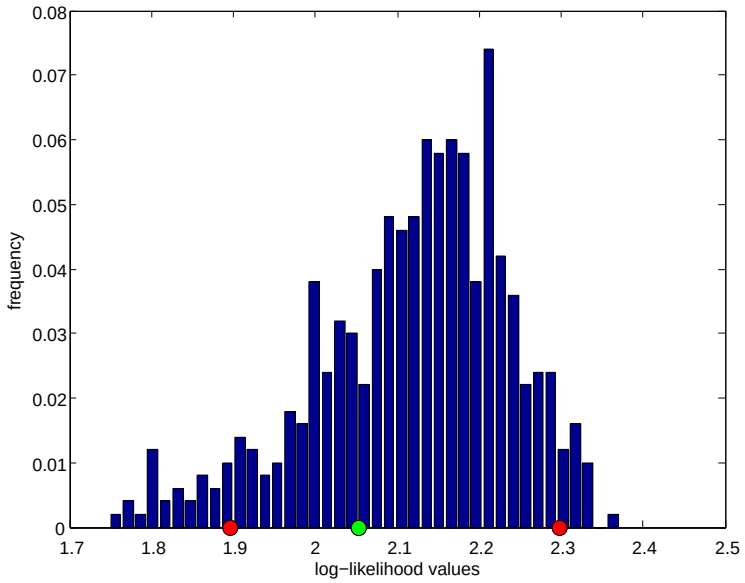


Figure 7.12: The empirical histogram average log-likelihood values for the best $EB(\mu, \lambda, a, b)$ exponential density on a bounded interval $[0.02, 0.16]$ with $\mu = 10 \times 10^{-7}$ and $\bar{s}(\Theta) = 0.05$. The true average log-likelihood pf observed s_{win} lies inside the quantiles Q_- and Q_+ of the simulated average log likelihood computed from s_{win} values.

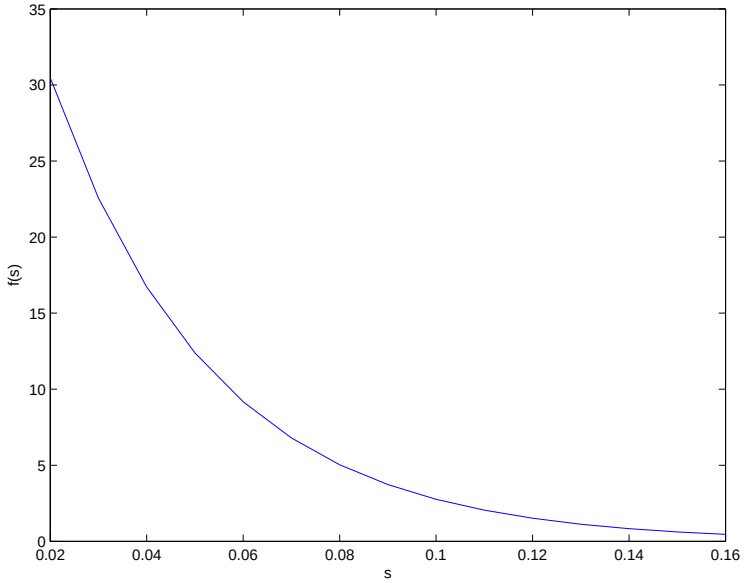


Figure 7.13: The density of selective advantages is plotted for exponential on the bounded interval [0.02,0.16] with mean selective advantage 0.05.

7.5 Study of Experiment 4

In this Section, we now study the 4th set of experiments that were started with initial genotype consisting of 3 mutations. It consists of mutations: Ribose, TopA, and SpoT. The initial genotype is the genotype with which the experiments are started and the selective advantage is computed with respect to this genotype, called the progenitor. The estimates of the selective advantages of the winner are computed by applying the non-linear least squares fitting as explained before in 6.5. We again explore the model with exponential density $EB(\mu, \lambda, a, b)$ on a bounded interval to assign the selective advantage of the mutants. We apply the tests $Test_{fix}$ and $Test_{swin}$ and display result for the best pair below. We explore systematically a wide range of values for μ and λ and hence for the mean $\bar{s}(\Theta)$. The estimates that we obtain for the selective advantages, s_{win} of winner applying non-linear least squares fitting (Section 6.5) are displayed in the table 7.8 below.

Table 7.8: The estimates of selective advantages and the fixation times as obtained for each population of Experiment 4.

Population	s_{win}	T_{fix}
Pop_1	0.05	125
Pop_2	0.15	38
Pop_3	0.04	162
Pop_4	0.08	74
Pop_5	0.17	65
Pop_6	0.08	192
Pop_7	0.14	186
Pop_8	0.2	131
Pop_9	0.06	134
Pop_{10}	0.06	80
Pop_{11}	0.11	56

The best model estimate for the mutation rate and the best quality fit was obtained for the model $EB(\mu, \lambda, a, b)$ with exponential density on the bounded interval $[0.03, 0.2]$, with parameter $\lambda = 30$. The estimate for the mutation rate is $\mu = 7 \times 10^{-7}$ and mean selective advantage $\bar{s}(\Theta) = 0.06$. We obtain a very good quality fit. The minimum $T_{min,\,obs} = 36$ and the maximum $T_{max,\,obs} = 194$. The probability $p(\Theta) = P(36 \leq T_{sim} \leq 194) = 0.78$ for $Test_{fix}$. The comparison of selective advantages, s_{win} of winner by computing the average log-likelihood is performed using $Test_{swin}$. The empirical histograms for the times to fixation are displayed below in figure 7.14. The empirical histograms for the selective advantages s_{win} of winner are also displayed (figure 7.15). We also display the empirical histograms for the average log-likelihoods (figure 7.16) as computed for the comparison ($Test_{swin}$). The plot for the exponential density with mean $\bar{s}(\Theta) = 0.06$ is also displayed (figure 7.17).

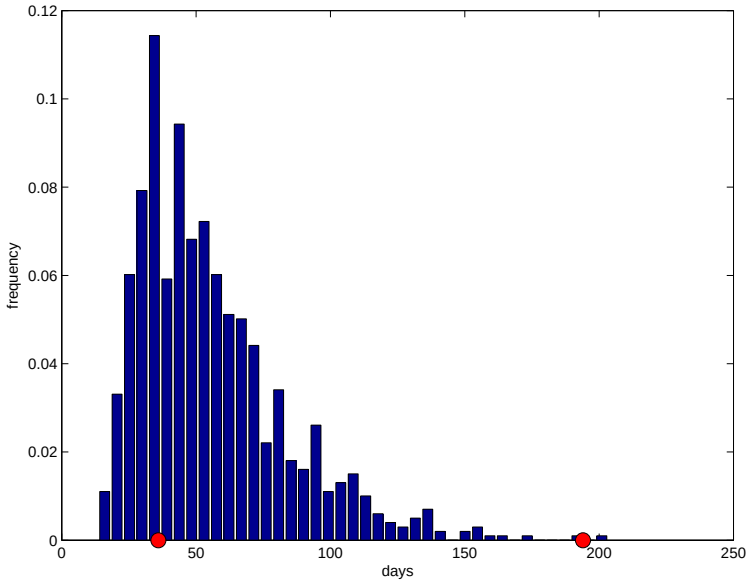


Figure 7.14: The empirical histogram of the fixation times for the best $EB(\mu, \lambda, a, b)$ exponential density on a bounded interval [0.03,0.2] with $\mu = 7 \times 10^{-7}$ and $\bar{s}(\Theta) = 0.06$. The red dots indicate the min and max of fixation times observed from the experimental data directly. The p-value $p(\Theta)^{11} = 6 \times 10^{-2}$.

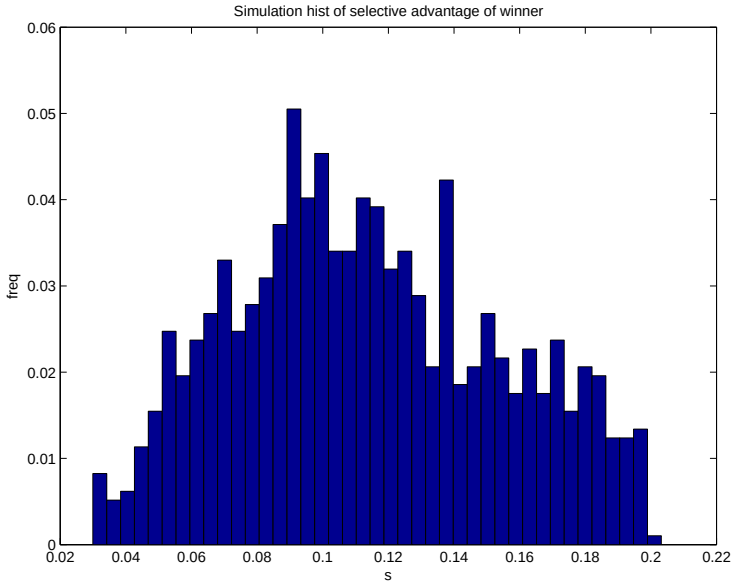


Figure 7.15: The empirical histogram of the selective advantage, s_{win} of the winner for the best $EB(\mu, \lambda, a, b)$ exponential density on a bounded interval [0.03,0.2] with $\mu = 7 \times 10^{-7}$ and $\bar{s}(\Theta) = 0.06$.

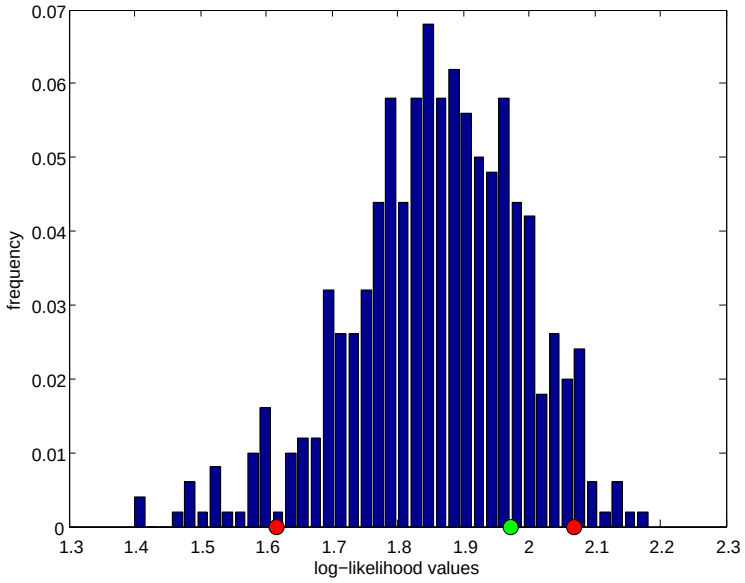


Figure 7.16: The empirical histogram average log-likelihood values for the best $EB(\mu, \lambda, a, b)$ exponential density on a bounded interval [0.03,0.2] with $\mu = 7 \times 10^{-7}$ and $\bar{s}(\Theta) = 0.06$. The true average log-likelihood pf observed s_{win} lies inside the quantiles Q_- and Q_+ of the simulated average log likelihood computed from s_{win} values.

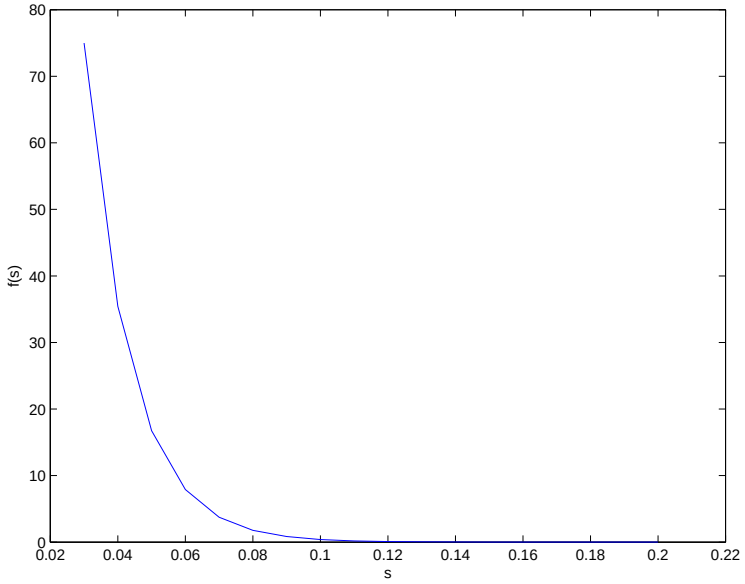


Figure 7.17: The density of selective advantages plotted for the best $EB(\mu, \lambda, a, b)$ exponential on bounded interval [0.03,0.2] with mean $\bar{s}(\Theta) = 0.06$.

7.6 Study of Experiment 5

In this Section, we study the 5th set of experiments that were started with initial genotype with 4 mutations. The initial genotype consisted of the 5 kinds of mutations, namely, Ribose + TopA + SpoT + glmus. The selective advantages of the winner are computed with respect to the initial genotype called the progenitor. These estimates are computed as explained above by applying the non-linear least squares fitting (Section 6.5) to the experimental population trajectories. We explore the model where the selective advantages of mutants are assigned based on $EB(\mu, \lambda, a, b)$ an exponential density on a bounded interval. We systematically explore a wide range of values for μ and λ. The estimated values for the selective advantage using the non-linear least squares fitting (6.5) of the data, and the times it takes for the mutation to reach fixation are given below in table 7.9:

Table 7.9: The estimates of selective advantages, s_{win} and the fixation times as obtained for each population of Experiment 5.

Population	s_{win}	T_{fix}
Pop_1	0.08	87
Pop_2	0.02	116
Pop_3	0.03	80
Pop_4	0.14	45
Pop_5	0.07	80
Pop_6	0.05	109
Pop_7	0.03	80
Pop_8	0.02	131
Pop_9	0.04	147
Pop_{10}	0.02	116
Pop_{11}	0.05	58

The best quality fit model and the best estimate for the mutation rate was obtained for $EB(\mu, \lambda, a, b)$ the exponential density on bounded interval $[0.01, 0.15]$ with parameter $\lambda = 50$ and the estimate of $\mu = 7 \times 10^{-7}$ and mean selective advantage is $\bar{s}(\Theta) = 0.03$. For this model, we obtain a very good quality fit. The $T_{min,\,obs} = 43$ and the $T_{max,\,obs} = 149$. The probability as obtained for $Test_{fix}$ is $p(\Theta) = P(43 \leq T_{sim} \leq 149) = 0.84$. The comparison for the selective advantages s_{win} for winner, using $Test_{swin}$ is performed. The empirical histograms for the times to fixation, and the empirical histograms for the selective advantages of the winner are displayed below (figures 7.18 and 7.19 respectively). We also display the empirical histograms for the average log-likelihood (figure 7.20). The density for selective advantages is plotted with mean $\bar{s} = 0.03$ (figure 7.21).

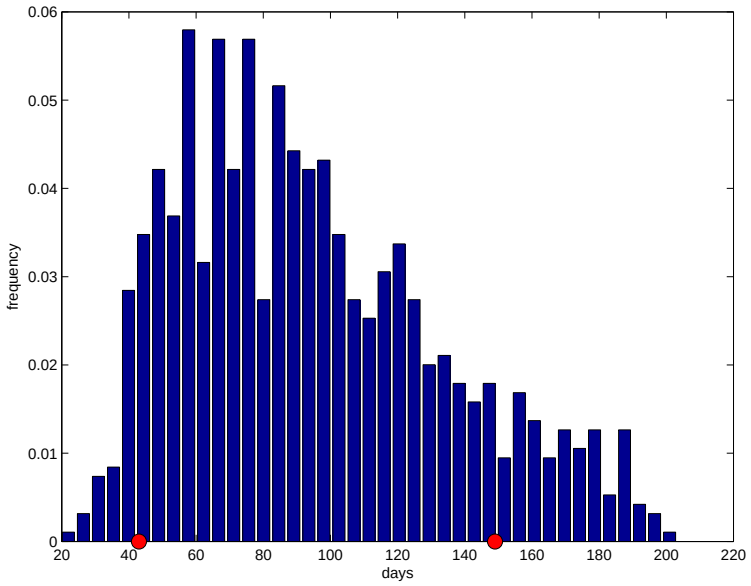


Figure 7.18: The empirical histogram of the fixation times for the best $EB(\mu, \lambda, a, b)$ exponential density on a bounded interval [0.01,0.15] with $\mu = 7 \times 10^{-7}$ and $\bar{s}(\Theta) = 0.03$. The red dots indicate the min and max of fixation times observed from the experimental data directly. We get $p(\Theta)^{11} = 0.15$.

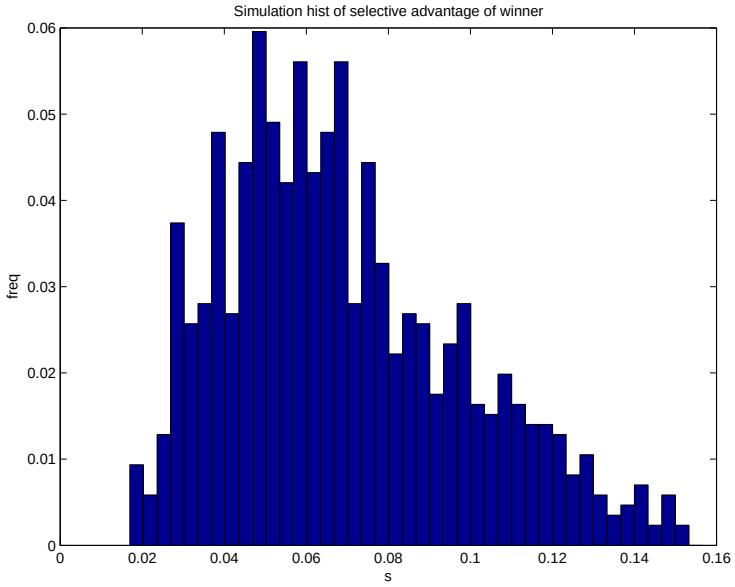


Figure 7.19: The empirical histogram of the selective advantage s_{win} of the winner for the best $EB(\mu, \lambda, a, b)$ exponential density on a bounded interval [0.01,0.15] with $\mu = 7 \times 10^{-7}$ and $\bar{s}(\Theta) = 0.03$.

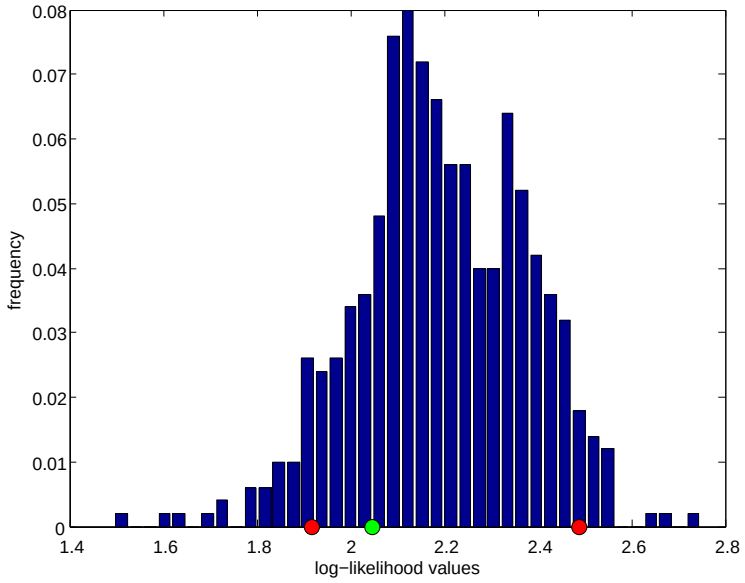


Figure 7.20: The empirical histogram average log-likelihood values for the best $EB(\mu, \lambda, a, b)$ exponential density on a bounded interval [0.01,0.15] with $\mu = 7 \times 10^{-7}$ and $\bar{s}(\Theta) = 0.03$. The true average log-likelihood pf observed s_{win} lies inside the quantiles Q_- and Q_+ of the simulated average log likelihood computed from s_{win} values.

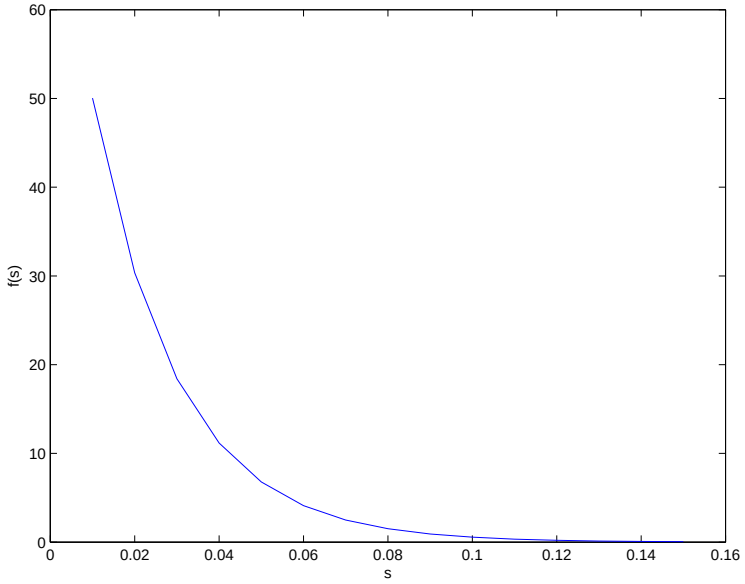


Figure 7.21: The density for selective advantages is plotted for $EB(\mu, \lambda, a, b)$, exponential on a bounded interval [0.01,0.15] with $\bar{s}(\Theta) = 0.03$.

7.7 Study of Experiment 6

In this Section, we study the last set of experimental data for the experiments that were started with initial genotype to have 5 mutations. The initial genotype consisted of the 5 kinds of mutation, namely, Ribose + TopA + SpoT + glmus + pykF. The selective advantages s_{win} of the winners are computed with respect to the initial genotype called the progenitor. These estimates are computed by non-linear least squares fitting algorithm as explained above (6.5). We explore systematically a range of values for the parameters μ and λ for the model $EB(\mu, \lambda, a, b)$ using an exponential density on the bounded interval with parameter λ. From the 11 experimental population trajectories of the winner, we can compute the time at which mutation goes to fixation. We display these values along with the estimated selective advantages using the non-linear least squares fitting (6.5) for winner for each population in Table 7.10.

Table 7.10: The estimates of selective advantages and the fixation times as obtained for each population of Experiment 6.

Population	s_{win}	T_{fix}
Pop_1	0.05	77
Pop_2	0.05	200
Pop_3	0.02	200
Pop_4	0.02	159
Pop_5	0.03	80
Pop_6	0.07	106
Pop_7	0.06	200
Pop_8	0.04	143
Pop_9	0.02	200
Pop_{10}	0.04	143
Pop_{11}	0.02	189

The best quality of fit to the model is obtained when we consider $EB(\mu, \lambda, a, b)$, the exponential on a bounded interval [0.01, 0.08] and with parameter $\lambda = 80$ for $\mu = 1 \times 10^{-7}$ and mean selective advantage $\bar{s}(\Theta) = 0.02$. We obtain a good quality fit for this model as well. From above table, we see that $T_{min,\,obs} = 75$ and $T_{max,\,obs} = 200$. The probability from $Test_{fix}$ is $p(\Theta) = P(75 \leq T_{sim} \leq 200) = 0.92$. The comparison of the selective advantages using the average log-likelihood using $Test_{swin}$ is performed. The empirical histograms for the times to fixation and the empirical histograms for the selective advantages of winner are displayed (figures 7.22 and 7.23 respectively). The empirical histogram for the average log-likelihood values are also displayed (figure 7.24). The plot for the exponential density function on the bounded interval with mean $\tilde{s} = 0.02$ is also displayed (figure 7.25).

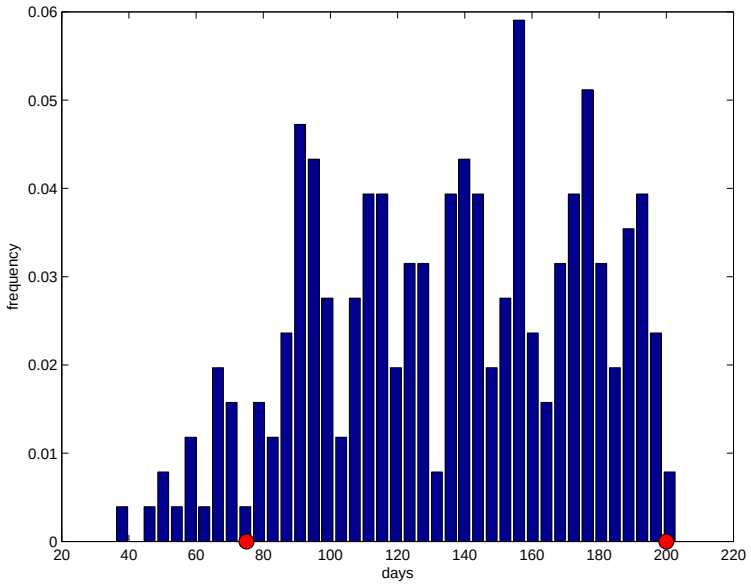


Figure 7.22: The empirical histogram of the fixation times for the best $EB(\mu, \lambda, a, b)$ exponential density on a bounded interval [0.01,0.08] with $\mu = 1 \times 10^{-7}$ and mean selective advantage $\bar{s}(\Theta) = 0.02$. The red dots indicate $T_{min,obs}$ and $T_{max,obs}$ of fixation times observed from the experimental data directly. The p-value $p(\Theta)^{11} = 0.4$.

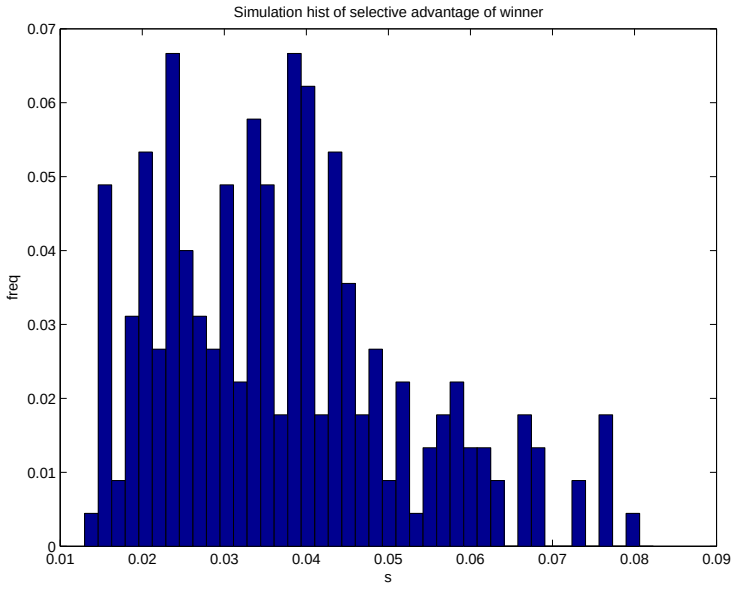


Figure 7.23: The empirical histogram of the selective advantages s_{win} of the winner for the best $EB(\mu, \lambda, a, b)$ exponential density on a bounded interval [0.01,0.08] with $\mu = 1 \times 10^{-7}$ and mean selective advantage 0.02.

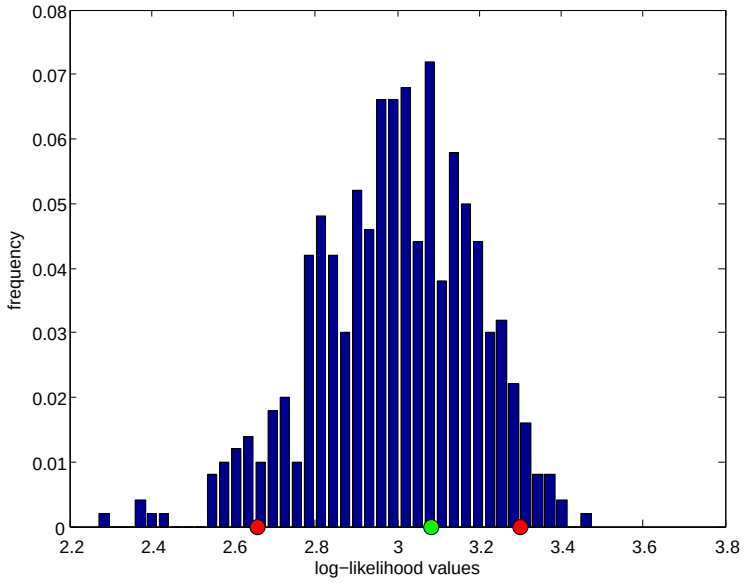


Figure 7.24: The empirical histogram average log-likelihood values for the best $EB(\mu, \lambda, a, b)$ exponential density on a bounded interval [0.01,0.08] with $\mu = 1 \times 10^{-7}$ and $\bar{s}(\Theta) = 0.02$. The true average log-likelihood pf observed s_{win} lies inside the quantiles Q_- and Q_+ of the simulated average log likelihood computed from s_{win} values.

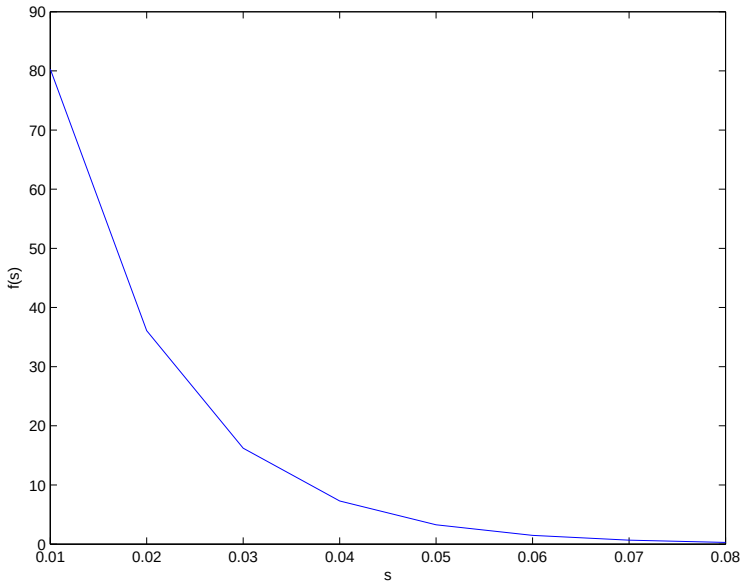


Figure 7.25: The density of selective advantages plotted for exponential on the bounded interval [0.01,0.08] with mean $\bar{s} = 0.02$.

7.8 Summary: Estimators and Accuracy

We have developed estimates for the mutation rate μ, as well as estimates for the mean $\bar{s}(\Theta)$ for the bounded exponential mean selective coefficient, for each of the experimental data set as obtained from T. Cooper's laboratory for the evolution of the population for the bacteria *Escherechia coli*. These estimates as above are obtained by using a maximum likelihood technique, where we first maximize over the probability for the range of fixation times ($Test_{fix}$), and then using the log-likelihood approach to compare the estimates for the selective advantages s_{win} of the winner using test $Test_{win}$. We summarize these results for all the experimental datasets, ranging from genotypes with no mutation, to genotypes with five different mutations.

To compute the accuracy of these estimators $\hat{\mu}$ and $\hat{\bar{s}}$, for each experiment, treat $\hat{\mu}$ and $\hat{\lambda}$ as the true unknown values for μ and λ. Thus for each experiment, let $\hat{\mu} = \mu_0$ and $\hat{\lambda} = \lambda_0$ be the true values. We then generate 100 estimates of μ and λ over the respective grids for μ and λ for each of the experiment (as mentioned in the above Sections). The mean selective advantage $\bar{s}$ can be computed from λ accordingly (chapter 6). We generate empirical distributions for estimates $\hat{\mu}$ and $\hat{\bar{s}}$ using the model $EB(\mu, \lambda, a, b)$ where the selective advantage of mutants are randomly sampled from exponential distribution on a bounded interval $[a, b]$. For each experiment, the interval $[a, b]$ for the model $EB(\mu, \lambda, a, b)$ is chosen such that the this interval contains the estimated selective advantages s_{win} of the winner (as computed using 6.5) for that

experiment. The mean $\bar{s}$ is then given by

$$\bar{s} \;=\; \frac{1}{\lambda} + \frac{ae^{-\lambda a} - be^{-\lambda b}}{e^{-\lambda a} - e^{-\lambda b}}.$$

Table 7.11 below displays the estimates $\hat{\mu}$ of the mutation rate and $\hat{\bar{s}}$ of mean selective advantage. The accuracy of these estimators can then be studied by computing the square root of the mean square error (σ), as follows:

$$\sigma_{\hat{\mu}} = \sqrt{\frac{1}{100}\sum_{i=1}^{100}(\hat{\mu}_i - \mu_0)^2} \;\; and \;\; \sigma_{\hat{\bar{s}}} = \sqrt{\frac{1}{100}\sum_{i=1}^{100}(\hat{\bar{s}}_i - \bar{s}_0)^2}$$

Table 7.11: The estimates for the mutation rate μ and mean selective advantage, as obtained for different experimental data studied above with multiple mutations model.

Experiment	$\hat{\mu}$	$\sigma_{\hat{\mu}}$	$\hat{\bar{s}}$	$\sigma_{\hat{\bar{s}}}$
Experiment 1	$10 * 10^{-7}$	1×10^{-7}	0.08	0.002
Experiment 2	$9 * 10^{-7}$	1.3×10^{-7}	0.07	0.003
Experiment 3	$10 * 10^{-7}$	1.6×10^{-7}	0.05	0.003
Experiment 4	$7 * 10^{-7}$	1.8×10^{-7}	0.06	0.005
Experiment 5	$7 * 10^{-7}$	2.3×10^{-7}	0.03	0.003
Experiment 6	$1 * 10^{-7}$	1.3×10^{-7}	0.02	0.001

For each of the empirical distributions of the estimators $\hat{\mu}$ and $\hat{\bar{s}}$, we compute a range of quantiles such that the 90% of the data lies within that range. Table 7.12 below displays these values.

Table 7.12: The 90% quantile range for the estimators $\hat{\mu}$ and $\hat{\bar{s}}$

Experiment	quantile range for $\hat{\mu}$	quantile range for $\hat{\bar{s}}$
Experiment 1	$[8 \times 10^{-7}\ ,\ 10^{-6}]$	[0.079 , 0.084]
Experiment 2	$[6.5 \times 10^{-7}\ ,\ 11 \times 10^{-7}]$	[0.062 , 0.069]
Experiment 3	$[8 \times 10^{-7}\ ,\ 12 \times 10^{-7}]$	[0.048 , 0.056]
Experiment 4	$[6 \times 10^{-7}\ ,\ 10^{-6}]$	[0.055 , 0.068]
Experiment 5	$[8 \times 10^{-7}\ ,\ 10^{-6}]$	[0.027 , 0.035]
Experiment 6	$[0.5 \times 10^{-7}\ ,\ 3 \times 10^{-7}]$	[0.022 , 0.024]

CHAPTER 8

Multiple Mutations: HK Experiments

In this chapter, we present similar bacterial evolution experiments and associated models studied by Hegreness et al. (2006) [23]. These (call them HK experiments) use an experimental setup similar to the TC experiments, to study the evolution of *E. coli* bacterial populations. The parameter values for the HK experiments, as mentioned earlier in chapter 3, are different from those for the TC experiments. Table 8.1 summarizes all the parameter values for both the TC and HK experiments. We first give a brief description of the simulation model for the HK experiments, and present their method of estimation for the mean selective advantage. We then

provide the accuracy for the estimators of mutation rate and mean selective advantage for HK estimation techniques as well the estimation technique developed in this thesis (chapter 7).

Table 8.1: Parameter values for the TC and HK experiments

Parameter	TC experiment	HK experiment
Experimental:		
$\mathcal{N}$	11	72
N_0	5×10^4	2.5×10^5
N_{sat}	10^7	8.25×10^8
D	200	3300
Model:		
τ	50	12
F	1.11	1.18
s	[0.01, 0.2]	[0.01, 0.2]
μ	$[2 \times 10^{-8}, 10^{-6}]$	$[2 \times 10^{-8}, 10^{-6}]$

8.1 Simulation Model

For the HK experimental setup, as before, the experiments begin with an initial genotype with $\mathcal{N}$ replicate populations, where $\mathcal{N} = 72$ is larger than the TC experiments, for which $\mathcal{N} = 11$. The number of initial *E.coli* bacterial cells in the beginning of the experiments is $N_0 = 2.5 \times 10^5$. They are evenly divided into "yellow" and "cyan" cells (tagged by neutral color markers). These $\mathcal{N}$ replicate populations, as before for the TC experiments, grow in parallel and are allowed to grow freely every day until the exhaustion of the nutrients, which occurs when the population reaches saturation size $N_{sat} = 8.25 \times 10^8$ cells. At the end of each daily growth phase, the population

undergoes a dilution with a dilution factor $D = \frac{N_{sat}}{N_0} \approx 3300$, i.e, N_0 cells are extracted from the current saturated population well and transferred to fresh medium, and allowed to grow again until population size N_{sat}. These {growth + dilution} cycles are performed daily. The counts for the numbers of the two marker cells are recorded at the end of each growth phase. These measurements are taken when the cells are in stationary phase. A plate reader, Victor III Perkin Elmer, recording the fluorescence readings is used to record numbers for the two sub-populations. This will add an error on the measurements. This is equivalent to the complementary second sub-sampling in the TC experiments, with higher accuracy.

We perform similar, multiple mutation simulations as described in chapter 6, by replacing the parameter values for the HK experiments. The counts for the number of yellow and cyan markers are recorded at the end of each growth phase. The simulations run until color fixation occurs. If the fixation has not occurred for 39 days, the simulation is stopped.

The selective advantage of the new emerging beneficial mutation, for this multiple mutation model $\mathcal{M}(\bar{s}, \mu)$, is sampled from exponential distribution with mean selective advantage, given by $\bar{s}$. Thus if a mutation occurs from the ancestor cells, then the selective advantage of that mutant with respect to the ancestor, is sampled from an exponential distribution with mean selective advantage $\bar{s}$. However, if the new mutation arises from the previously born mutant with selective advantage s_{old}, then the selective advantage of this new arising mutation is given as below:

$$s_{new} = s_{old} + \triangle s$$

where $\triangle s$ is sampled from exponential distribution with mean $\bar{s}$. Thus, the selective advantage of the new arising mutation from a previously born mutant is the sum of the selective advantage of the mutant from which the new mutation arises incremented by an randomly sampled $\triangle s$ picked from exponential distribution with mean $\bar{s}$. This method of selecting selective advantages for newly arising mutations, is equivalent to the conditional method (as studied above in chapter 7) for the exponential density on $[0, \infty]$.

The ratios of the counts for the size of the yellow and cyan sub-populations at the end of the growth day are recorded. Let $N_Y(d)$ and $N_C(d)$ denote the measurements for the yellow and cyan markers respectively, at the end of the growth day d.

8.2 Simulation Data Base

We consider a grid of pairs $(\mu, \bar{s})$ as follows: 10 values of mutation rate, $\mu \in \{10^{-7}, 2 \times 10^{-7}, \ldots, 10^{-6}\}$ and 20 equally spaced values of mean selective advantage $\bar{s} \in [0.01, 0.2]$. For each pair $(\mu, \bar{s})$, we ran the simulations of the multiple mutations HK model described above in Section 8.1, we generate 500 random evolution process trajectories. Thus generating 500 random evolution process trajectories for 200 pairs, to consist of our simulation data base.

8.3 HK Estimation: Fitting the Initial Divergence of $g(t)$

The two sub-populations start with the same initial size and initial genotype. Mutations occur in both sub-populations and causes initial fluctuations in the curve $g(t) = \log \frac{p(t)}{1-p(t)}$, until a mutant with stronger selective advantage emerges causing the curve to deviate from the almost flat line. Thus, beneficial mutations cause the curve $g(t)$ to deviate from its starting value. We describe here the fitting of the marker ratio curves concentrating on the initial phase of the experiment, thus fitting this initial divergence of the marker ratio curves for the HK experiments. The sizes of the yellow and cyan populations are recorded as explained above. For each one of the 72 populations monitored, the following statistical model is used to fit the initial divergence of the marker ratio curves:

$$g(t) = \log\left(1 + \frac{1}{2}\exp\left(\hat{\alpha}(t - \hat{\kappa})\right)\right) + c_t + \epsilon_t \tag{8.1}$$

where $\hat{\alpha}$ and $\hat{\kappa}$ are the parameters of the growth curve specific to each well. Here, $\hat{\kappa}$ denotes the first time when a significant deviation from the flat line in the curve $g(t)$ occurs, and $\hat{\alpha}$ is the slope of this deviation. The c_t denotes the daily biases observed in the data due to the technique used for recording the sizes of the two sub-populations and $\epsilon_t \sim N(0, \sigma^2)$ are independent and describe the uncertainty in each measurement.

The time t in equation (8.1), denotes the number of generations, i.e, $t = 12d$ where d is the day at which the measurement was taken. Since these experiments are performed for approximately 450 generations, we have days, $d = 1, \ldots, 39$.

A non-linear least squares (NLLS) regression is applied to fit the curve $g(t)$ by equation (8.1). The residuals of the fit are checked for conditions C(1) and C(2) below. The curve $g(t)$ is fit until the last time the conditions C(1) and C(2) are both satisfied, and thus extracting $\hat{\alpha}$ and $\hat{\kappa}$ from this fitting. The authors (Hegreness et al. (2006) [23]) make use of the Lilliefors test to test the null hypothesis that the data comes from a normally distributed population (when the null hypothesis do not specify the mean and variance of the normal distribution). The measurements in each well are included in the fit up to a time t which is the latest time satisfying the following two conditions:

C(1) Goodness of fit to the data: Lilliefors test at 5% significance level is used to accept or reject the hypothesis.

C(2) The standard deviation of the residuals of the fit does not exceed 0.15.

Estimating these times $\hat{\kappa}$ and slopes $\hat{\alpha}$ for each of the 72 marker-ratio trajectories, thus generates virtual values of $\hat{\alpha}$ and $\hat{\kappa}$. Call these 72 virtual values of $\hat{\alpha}$ and $\hat{\kappa}$, the values obtained from the *E. coli* experimental data. Figure 8.1 displays an example of the curve $g(t)$ for the HK simulations, when the model $\mathcal{M}(0.02, 5 \times 10^{-7})$ is used (see 8.1). We plot the curve $g(t)$, and the fitted curve $\log\left(1 + 0.5 \exp\left(\hat{\alpha}(t - \hat{\kappa})\right)\right)$ to estimate the first significant deviation time $\hat{\kappa} = 200$ generations (i.e 16 days), slope $\hat{\alpha} = 0.07$ of this deviation. This fitting is performed on $g(t)$ up to time t, such that the conditions C(1) and C(2) are satisfied. The conditions C(1) and C(2) are satisfied up to generation 276 (i.e up to day 23) for this trajectory.

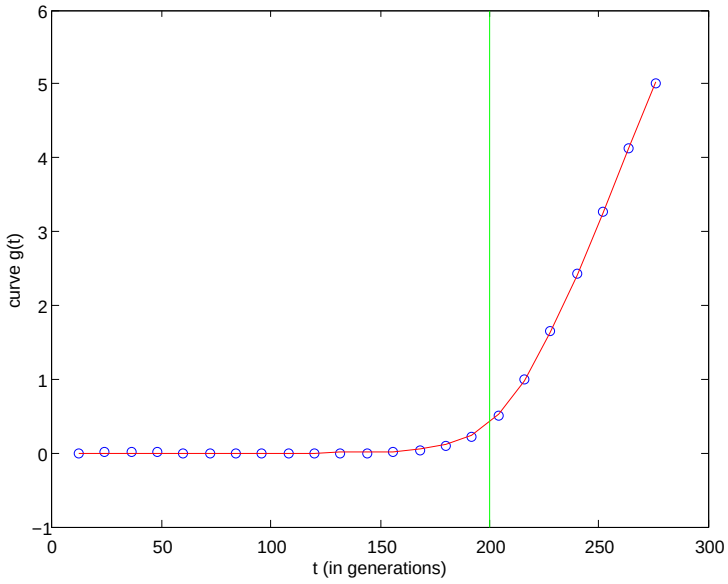


Figure 8.1: The blue dots display the curve $g(t)$, versus time in generations. The best fit curve ($\hat{\alpha} = 0.07$ and $\hat{\kappa} = 200$ generations) is displayed in red (solid line) and $\hat{\kappa} = 200$ generations is indicated by the green line. For this example, $\bar{s} = 0.02$ and $\mu = 5 \times 10^{-7}$.

8.4 Evaluating the Performance of Estimators

For each pair $(\bar{s}, \mu)$ in the simulation data base above in Section 8.2, we have generated 500 random evolution process trajectories. We apply the fitting to the curve $g(t)$, as explained in Section 8.3 to each of the 500 process trajectories for each of the 200 pairs. This generates empirical histograms for $\hat{\alpha}$ and $\hat{\kappa}$ from simulation trajectories. We use the simulated data (described in Section 8.1 and 8.2), generated by our simulations of the HK model using the HK experiment parameters. We separately apply the HK estimation technique described in Section 8.3, on this simulated data. This generates the above empirical histograms of $\hat{\alpha}$ and $\hat{\kappa}$ for each pair $(\bar{s}, \mu)$. To evaluate the performance of the estimators, the 72 virtual values of $\hat{\alpha}$ and $\hat{\kappa}$ obtained

from experimental data are compared with the 500 empirical samples of $\hat{\alpha}$ and $\hat{\kappa}$ for each pair $(\bar{s}, \mu)$. Kolmogorov-Smirnov (KS) test is applied at 2.5% significance level to compare the 72 virtual values of $\hat{\alpha}$ with the empirical histograms of $\hat{\alpha}$ for all pairs $(\bar{s}, \mu)$, with the null hypothesis that the 72 virtual values of $\hat{\alpha}$ from experimental data and the empirical histograms of $\hat{\alpha}$ come from the same distribution. This generates pairs $(\bar{s}, \mu)$ for which the null hypothesis is accepted at 2.5% significance level, for α. Similarly, KS test is applied at 2.5% significance level to compare the 72 virtual values of $\hat{\kappa}$ from experimental data with the empirical histograms of $\hat{\kappa}$ for all pairs $(\bar{s}, \mu)$, with the null hypothesis that the 72 virtual values of $\hat{\kappa}$ from experimental data and the empirical histograms of $\hat{\kappa}$ come from the same distribution. This generates pairs $(\bar{s}, \mu)$ for which the null hypothesis is accepted at 2.5% significance level, for $\hat{\kappa}$. This generates estimates for pairs $(\bar{s}, \mu)$ where the null hypothesis is not rejected for both $\hat{\alpha}$ and $\hat{\kappa}$. Since these two distributions of $\hat{\alpha}$ and $\hat{\kappa}$ are not independent, these common pairs of $(\bar{s}, \mu)$ where the null hypothesis is not rejected for both $\hat{\alpha}$ and $\hat{\kappa}$ represents at least a 95% confidence regions for the estimators $(\hat{\bar{s}}, \hat{\mu})$ of mutation rate, μ and mean selective advantage, $\bar{s}$.

8.5 Application to Virtual Experimental Values

Since we do not have access to the actual HK experimental data, we generate virtual experimental values. We apply the HK estimation method for the evaluation of the mean selective advantage $\hat{\bar{s}}$ and the mutation occurrence $\hat{\mu}$. We compare the HK method to the statistical techniques we have introduced to select a multiple mutations model among the 3 categories of models presented above in chapter 7 . We then provide the accuracy of these estimators for the HK estimation technique, and the estimation technique developed for the multiple mutation models in chapter 7 .

8.5.1 HK Estimation Technique

We extract 72 virtual values of $\hat{\alpha}$ and $\hat{\kappa}$ using the technique described in 8.3 by simulating the true model $\mathcal{M}(s_0, \mu_0)$ where $s_0 = 0.08$ and $\mu_0 = 7 \times 10^{-7}$. We consider these 72 virtual values as the estimated values for $\hat{\alpha}$ and $\hat{\kappa}$ derived from the experimental data. We then apply the estimation technique for HK as described above (8.2, 8.3 and 8.4). This generates empirical histograms of $\hat{\alpha}$ and $\hat{\kappa}$ for each pair $(\bar{s}, \mu)$. Comparing the virtual values with empirical histograms of $\hat{\alpha}$ and $\hat{\kappa}$ for all the 200 pairs of $(\bar{s}, \mu)$ (as described in 8.4), we generate the HK confidence regions at 95% confidence level. Figure 8.2 displays this region using the HK estimation technique (Section 8.3 and 8.4). Note that the HK estimation technique provides the pairs of $(\bar{s}, \mu)$ representing the 95% confidence region, and we compute the estimate $(\hat{\bar{s}}, \hat{\mu})$ as the barycenter of these pairs.

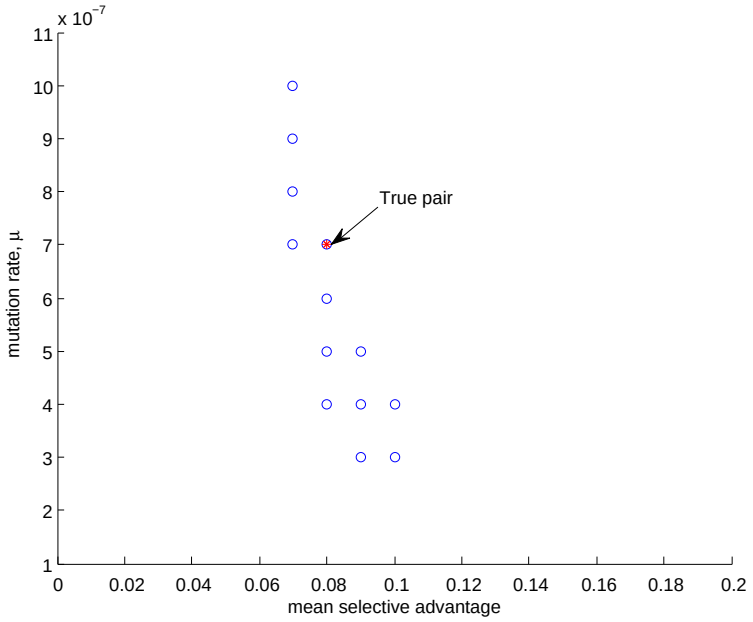


Figure 8.2: The confidence region obtained on applying the HK estimation for true $(s_0, \mu_0) = (0.08\ ,\ 7 \times 10^{-7})$. Points $(\bar{s}, \mu)$ indicate the pairs for which the null hypothesis (that 72 virtual values and 500 empirical histograms of $\hat{\alpha}$ and $\hat{\kappa}$ come from the same distribution) using KS test at 2.5% significance is not rejected.

We verified by the following algorithm, the HK- confidence regions at 95% confidence level:

1. Fix a true pair (s_0, μ_0).

2. Extract 72 virtual values of $\hat{\alpha}$ and $\hat{\kappa}$ representing the values of $\hat{\alpha}$ and $\hat{\kappa}$ as derived from experimental data.

3. Compare these 72 virtual values with the 500 empirical values of $\hat{\alpha}$ and $\hat{\kappa}$ for each pair $(\bar{s}, \mu)$ (as above in 8.4). This generates a region of pairs $(\bar{s}, \mu)$ for which the null hypothesis is not rejected by KS test at 2.5% significance. Example of one such region is displayed in figure 8.2.

4. Repeat steps 2 and 3 above 100 times generates 100 such regions, and each time check whether the true pair $(s_0, \mu_0) \in$ region or not.

We see that 96 out of 100 runs above, the true pair $(s_0, \mu_0) \in$ region, thus verifying that comparison as above (in 8.4) generates HK-confidence region at least at 95% confidence level.

To compute an estimate $(\hat{\bar{s}}, \hat{\mu})$ from the pairs $(\bar{s}, \mu) \in$ region, we consider the barycenter of that region. This process generates empirical histograms for estimates $\hat{\bar{s}}$ and $\hat{\mu}$. For the region displayed in figure 8.2, the estimate $(\hat{\bar{s}}, \hat{\mu}) = (0.08, 6 \times 10^{-7})$. Since, the true pair $(s_0, \mu_0) = (0.08, 7 \times 10^{-7})$ is known, we compute the accuracy of the estimators $\hat{\bar{s}}$ and $\hat{\mu}$ by computing the square root of the mean square error as below:

$$msq_{\hat{\bar{s}}} = \sqrt{\frac{1}{100}\sum_{i=1}^{100}(\hat{\bar{s}}_i - s_0)^2} \;\; and \;\; msq_{\hat{\mu}} = \sqrt{\frac{1}{100}\sum_{i=1}^{100}(\hat{\mu}_i - \mu_0)^2}$$

We obtain $msq_{\hat{\bar{s}}} = 0.01$ and $msq_{\hat{\mu}} = 2 \times 10^{-7}$.

8.5.2 Thesis Estimation Technique

In chapter 7, we developed estimators $\hat{\bar{s}}$ and $\hat{\mu}$ for multiple mutation models using the maximum likelihood approach as in $Test_{fix}$ and $Test_{swin}$. We simulate the model with exponential density on $[0, \infty]$ for randomly selecting selective advantages of mutants, with the HK experimental and model parameters (8.1). We thus obtain as above 500 replications of random evolutionary trajectories for each pair $(\bar{s}, \mu)$. We then apply the tests $Test_{fix}$ and $Test_{swin}$ to obtain estimates $\hat{\bar{s}}$ and $\hat{\mu}$.

We randomly extract 72 virtual replications from true $(s_0, \mu_0) = (0.08, 7 \times 10^{-7})$.

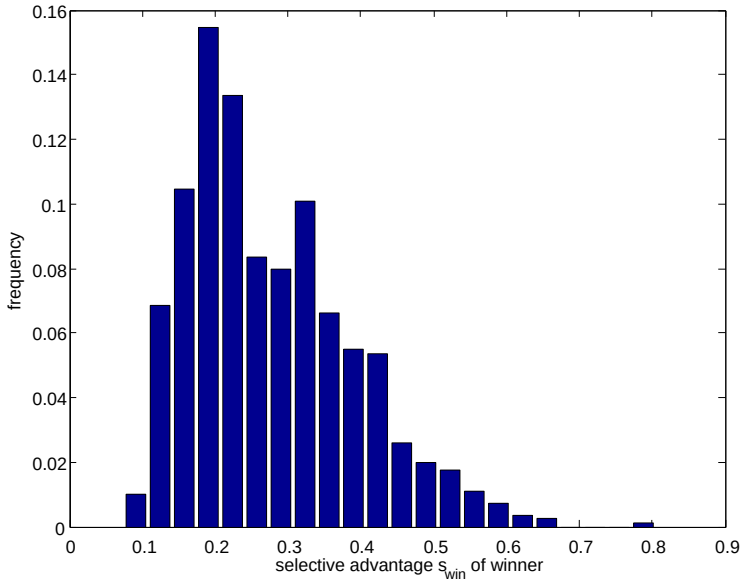


Figure 8.3: The selective advantages s_{win} of winner as obtained from simulations, for estimates of HK ($\hat{\bar{s}} = 0.05$ and $\hat{\mu} = 10^{-6}$).

Consider these to be the experimental data from HK experiments. Repeating this 100 times, provides 100 sets for experimental data of 72 replications. We apply the estimation technique (as in chapter 7), to maximize the probabilities obtained from $Test_{fix}$ and $Test_{swin}$, and obtain 100 estimates for $\hat{\bar{s}}$ and $\hat{\mu}$. The accuracy can then be computed as above by calculating the square root of the mean square error. We obtain $msq_{\hat{\bar{s}}} = 0.02$ and $msq_{\hat{\mu}} = 2 \times 10^{-7}$.

In conclusion, we can make the following comparison between our estimation method and that in the Hegreness et al. [23] paper.

1. First, if we compare the selective advantages of the winner from the experimental data obtained from TC's experiments with the selective advantages of the winner from the simulated data generated by using the exponential

distribution for picking the selective advantages, then the simulated selective advantages that we get for the winner are much larger than those seen in the experiments. Figures 7.1 and 8.3 display these examples.

Since we don't have access to the HK experimental data, we are generating their virtual values for the experimental data by following the scheme described above. For this reason there is a better match between the "experimental data" for HK and the simulated data for HK than we expect there to be. Even then, our accuracies are comparable to the accuracies obtained from their estimation method.

2. The parameter ranges that we are interested in are different from the parameter ranges for which they claim the exponential distribution is the correct model. For the parameter ranges of our interest, the model that they choose is the dirac delta distribution.

3. Finally, as of now, to compute the best fit model by the estimation technique, we have compared two histograms: for s_{win} and T_{fix}. A more detailed comparison may yield a better understanding of the strengths of each estimation method.

CHAPTER 9

Conclusions and Further Discussion

We have introduced and rigorously studied a detailed Poisson process model for the evolutionary dynamics of evolving populations of asexual *Escherichia coli* bacteria. The model formalizes two types of biological experiments (the TC and HK experiments), which evolve $\mathcal{N}$ replicate bacterial populations starting with initial individuals having identical genotypes. All initial cells are tagged by one of two neutral markers, red or white, which are transmitted by cell divisions. Each of these populations was submitted to a daily {growth + dilution} cycle before daily transfer to a new well. The TC experiments have a daily complementary dilution step , which involves extracting a few hundred cells from each new well in order to estimate the

current frequency of color markers by visual counts. The first daily dilution introduces what is known as genetic drift in the population and rare beneficial mutations may get lost through these daily dilutions.

We first study a stochastic model for the dynamics of asexual bacterial population evolving under the simplifying assumption that there is only one single type of irreversible mutation. We have focused on developing efficient estimators for two fundamental evolutionary parameters, the selective advantage, s, of newly arising adaptive mutations and the rate, μ, at which these mutations occur. We have developed new algorithms to estimate these parameters with no bias and to analyze the precision of these estimates and their sensitivity to the number $\mathcal{N}$ of experimental replications.

We have also studied the asymptotic behavior of these estimators as $\mathcal{N}$ becomes large and implemented simulations to study the estimation accuracy for realistic much smaller numbers of experimental populations such as $\mathcal{N} = 11$ or $\mathcal{N} = 72$.

Our estimator $\hat{s}$ of the selective advantage s is quite accurate, even for moderate values of $\mathcal{N}$. We find that $\hat{s}$ has no appreciable bias. The root mean squared error of estimation for $\hat{s}$ decreases from 0.02 for $\mathcal{N} = 1$ replicate population to 0.006 when replication is increased to $\mathcal{N} = 11$ as in the TC experiment, and to 0.002 when replication is increased to $\mathcal{N} = 72$ as in the HK experiment. The "ground truth values" of s measured experimentally for populations evolved in the TC experiment, were only available for one of the six TC experiments. Our estimators were able to accurately predict the "ground truth" values of s. This is important because it is likely that the $\mathcal{N}$ population replicates have different types of beneficial mutations.

This situation deviates from the single class of beneficial mutation considered in our first model. Evidently, our method is robust to this kind of deviation from our underlying simplification. The main reason for this is that smaller effect mutations are out-competed when they are at low frequency and thus do not reach frequencies high enough to influence the dynamics of neutral marker (Rozen et al. (2002) [51]). This explanation was explored in a previous paper, which demonstrated that similar color marker dynamics result even when very different distributions of underlying beneficial mutation effects are available to the evolving population (Hegreness et al. (2006) [23]).

We then explored the effect of the complementary sub-sampling dedicated to the experimental evolution of daily color marker frequencies, as was the case with the TC experiments. We have modified our estimators $\hat{s}$ and $\hat{\mu}$ to better handle this specific experimental feature which increases the errors of marker frequencies experimental evaluations. We compare the accuracies of these modified estimators to their accuracy when daily red and white marker frequencies are acquired with high precision. We have also studied the accuracy of our estimators $\hat{s}$ and $\hat{\mu}$ as the size N_{sub} of the second sub-sampling is increased. Note that this complementary sub-sampling does not modify the evolution of the population.

We find that the accuracy of the estimator $\hat{s}$ is reduced when color marker frequencies are evaluated by complementary sub samples of small sizes 300 to 400. The root mean squared error for $\hat{s}$ is inferior to 0.02 when we have $\mathcal{N} = 1$ replicate population and highly precise evaluations of daily color marker frequencies; this error increases to 0.03 when frequencies are evaluated by complementary sub-samples of

size $N_{sub} = 400$ (as was the case for the TC experiments). The error for the estimator $\hat{\nu}$ is less than 0.8 for $\mathcal{N} = 11$ and highly precise marker frequencies evaluations, while this error increases to 1.5 in the context of complementary sub-sampling of size $N_{sub} = 400$. We studied the asymptotic behavior of our estimators as N_{sub} becomes large. We show that the accuracy of our estimators increases as N_{sub} increases, and this enables us to recommend more realistic numbers for the size of N_{sub} to ensure adequate experimental accuracy.

Direct measurement of beneficial mutation rates is notoriously difficult, and is currently possible for only a limited number of strain-mutation combinations (Cooper et al. (2001) [8]). Thus, it was not possible to compare against ground truth our estimator $\hat{\mu}$ of mutation occurrence rates based on color marker dynamics. We have shown that our estimates of beneficial mutation rates are sensitive to the number $\mathcal{N}$ of population replications. For small values of $\mathcal{N}$, such as $\mathcal{N} = 11$, the error of estimation associated to our estimators of logarithmic mutation rate approaches 6% when there is no complementary second sub-sampling effect. However, for $\mathcal{N} \geq 30$, our estimates $\hat{\nu}$ of $\nu = \log \mu$ becomes much more accurate, with error of estimation less than 2% of the true value. Similar experiments designed to enable estimation of the underlying beneficial mutation rate, should take into account the fairly fast increase in accuracy due to higher experimental replication number $\mathcal{N}$.

We have then developed and studied parametric estimation for more complex models involving multiple mutations types. We applied new model fitting techniques to determine the best fit model for TC experimental data. Each one of these 6 TC experiments start with a distinct initial genotype and has $\mathcal{N} = 11$ replications.

We show that for the 6 best fit models associated to these 6 TC experiments, the density function for the random selective advantages of beneficial mutations can be efficiently modeled by an exponential density with a suitable parameter λ, restricted to a specific bounded interval [a,b]. To select and parametrize multiple mutations models which have the best fit to experimental data, we first compute, for a large finite family of model parameters, the simulated empirical histograms of the fixation times and of the selective advantage s_{win} of the winning mutant. We then develop statistical tests $Test_{fix}$ and $Test_{swin}$ based on precise log-likelihood comparisons between each one of the two preceding histograms and its corresponding counterpart histogram extracted from the experimental data.

We also compute the accuracy for the two estimators $\hat{\mu}$ and $\hat{\tilde{s}}$ derived from our model fitting techniques, for each of the six TC-experiments.

We have analyzed the previous multiple mutation models studied in (Hegreness et al. (2006) [23]) and (Barrick et al. (2010) [5]). We have extended their multiple mutations model as well as their parameter estimation techniques. To improve the precision of simulations for this type of process, we approximate the continuous growth phase by dividing it into 50 time intervals instead of 12 intervals in previous work; in particular, the key probability of successful bottleneck crossing, is proved to be much closer to its true values using our simulations techniques. We have developed efficient estimators of the rate and mean selective advantage of beneficial mutations based on the evolution of color frequency markers, and we have also provided detailed accuracy evaluations for these two estimators in the context of multiple mutations models.

Bibliography

[1] J.G. Arjan, M. De Visser, C.W. Zeyl, P.J. Gerrish, J.L. Blanchard, and R.E. Lenski. Diminishing returns from mutation supply rate in asexual populations. *Science*, 283(5400):404–406, January 1999.

[2] K.C. Atwood, L.K. Schneider, and F.J. Ryan. Periodic selection in *E. coli*. *Genetics*, 37:146–155, 1951.

[3] R. Azencott. Grandes deviations et applications. *Lecture notes in Mathematics*, 774:1–176, 1980.

[4] J.E. Barrick and Lenski R.E. Genome-wide mutational diversity in an evolving population of *Escherichia coli*. *Cold Springs Harbor Symposia on Quantitative Biology*, 74(18):119–129, 2009.

[5] J.E. Barrick, C.C. Strelioff, and R.E. Lenski. *Escherichia coli* rpob mutants have increased evolvability in proportion to their fitness defects. *Molecular Biol. Evol.*, 27:1338–1347, 2010.

[6] T. Bataillon, T. Zhang, and R. Kassen. Cost of adaptation and fitness effects of beneficial mutations in pseudomonas fluorescens. *Genetics*, 111, August 2011.

[7] V.S. Cooper and R.E. Lenski. The population genetics of ecological specialization in evolving *E. coli* populations. *Nature*, 407:736–739, 2000.

[8] 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*. *B. J. Bacteriol*, 9(183):2834–2841, 2001.

[9] C. Darwin. The origin of species by means of natural selection. *Publisher John Murray*, November 1859.

[10] J. De Visser and D.E. Rozen. Clonal interference and periodic selection of new beneficial mutations in *Escherichia coli*. *Genetics*, 172:2093–2100, 2006.

[11] M.M. Desai and D.S. Fisher. Beneficial mutation-selection balance and the effect of linkage on positive selection. *Genetics*, 176:1759–1798, 2007.

[12] M.M. Desai, D.S. Fisher, and A.W. Murray. The speed of evolution and maintenance of variation in asexual populations. *Current Biology*, 17:385–394, March 2007.

[13] W.J. Ewens. The probability of survival of a new mutant in a fluctuating environment. *Heredity*, 43:438–443, 1967.

[14] W.J. Ewens. Mathematical population genetics: I. theoretical introduction. *Springer, New York*, 2004.

[15] W. Feller. An introduction to probability theory and its applications. *John Wiley and Sons Inc*, 1968.

[16] R.A. Fisher. The distribution of gene ratios for rare mutations. *Contributions to mathematical statistics*, 50:205–220, May 1930.

[17] C.A. Fogle, J.L. Nagle, and M.M. Desai. Clonal interference, multiple mutations and adaptation in large asexual populations. *Genetics*, 180:2163–2173, 2008.

[18] P.J. Gerrish and R.E. Lenski. The fate of competing beneficial mutations in an asexual population. *Genetica*, 102(103):127–144, 1998.

[19] J.H. Gillespie. The causes of molecular evolution. *Oxford University Press*, 1991.

[20] D. Gresham, M. Desai, Tucker C.M., H.T. Jenq, and D.A. et al Pai. The repertoire and dynamics of evolutionary adaptations to controlled nutrient-limited environments in yeast. *PLoS Genetics*, 4(12):e1000303, December 2008.

[21] J.B.S. Haldane. A mathematical theory of natural and artificial selection part v: Selection and mutation. *Proceedings of Cam Phil Sc*, pages 838–844, July 1927.

[22] J.M. Heffernan and L.M. Wahl. The effects of genetic drift in experimental evolution. *Theoretical Population Biology*, 62:349–356, July 2002.

[23] M. Hegreness, N. Shoresh, D. Hartl, and R. Kishony. An equivalence principle for the incorporation of favorable mutations in asexual populations. *Science*, 311:1615–1617, 2006.

[24] R.B. Helling, C.N. Vargas, and J. Adams. Evolution of *E. coli* during growth in a constant environment. *Genetics*, 116:349–358, 1987.

[25] M. Imhof and C. Schlotterer. Fitness effects of advantageous mutations in evolving *Escherichia coli* populations. *Proc. Nat. Acad. Sci. USA*, 98(3):1113–1117, Jan 2001.

[26] S.B. Joseph and D.W. Hall. Spontaneous mutations in diploid saccharomyces cerevisiae: More beneficial than expected. *Genetics*, 168:1817–1825, December 2004.

[27] K.C. Kao and G. Sherlock. Molecular characterization of clonal interference during adaptive evolution in asexual populations of *Saccharomyces cerevisiae*. *Nat. Genetics*, 40(12):1499–1504, December 2008.

[28] R. Kassen and T. Batailon. Distribution of fitness effects among beneficial mutations before selection in experimental populations of bacteria. *Nature Genetics*, 38(4):484–488, April 2006.

[29] M. Kimura. Some problems of stochastic processes in genetics. *Ann. Math. Stat.*, 28:882–901, 1957.

[30] M. Kimura. On the probability of fixation of mutant genes in a population. *Genetics*, 47(6):713–719, 1962.

[31] M. Kimura. Model of effectively neutral mutations in which selective constraint is incorporated. *Proceedings of the National Academy of Sciences of the United States of America*, 76(7):3440–3444, July 1979.

[32] M. Kimura. The neutral theory of molecular evolution. *Cambridge Univ Press*, 1983.

[33] Leslie A. Lamport. The gnats and gnus document preparation system. *G Animals Journal*, 41(7):73+, july 1986.

[34] R.E. Lenski. Experimental studies of pleiotropy and epistasis in *Escherichia coli* variation in competitive fitness among mutants resistant to virus t4. *Evolution*, 42:425–433, 1988.

[35] 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. *American Naturalist*, 138(6):1315–1341, Dec 1991.

[36] R.E lenski and M. Travisano. Dynamics of adaptation and diversification: a 10000 generation experiment with bacterial populations. *Proc Natl Acad Sci USA*, 91:6808–6814, 1994.

[37] B.R Levin, V. Perrot, and N. Walker. Compensatory mutations, antibiotic resistance and the population genetics of adaptative evolution in bacteria. *Genetics*, 154:985–997, 2000.

[38] B.R. Levin, F.M. Stewart, and L. Chao. Resource limited growth, competition, and predation: a model and experimental studies with bacteria and bacteriophage. *American Naturalist*, 111:3–24, 1977.

[39] R.C. MacLean and A. Buckling. The distribution of fitness effects of beneficial mutations in pseudomonas aeruginosa. *PLoS Genet.*, 5(3), March 2009.

[40] M.J. McDonald, T.F. Cooper, H.J.E. Beaumont, and P.B. Rainey. The distribution of fitness effects of new beneficial mutations in pseudomonas aeruginosa. *Biol. Lett*, 7(1):98–100, Feb 2011.

[41] R. Miralles, P.J. Gerrish, A. Moya, and S.F. Elena. Clonal interference and the evolution of *RNA* viruses. *Science*, 285:1745–1747, 1999.

[42] H.J. Muller. Further studies on the nature and causes of gene mutations. *Proceedings of Sixth International Congress of Genetics*, pages 213–255, 1932.

[43] T. Ohta. Extension of the neutral mutation drift hypothesis. *Proceedings of the Second Taniguchi International Symposium on Biophysics*, pages 148–167, 1977.

[44] L. Perfeito, L. Fernandes, C. Mota, and I. Gordo. Adaptive mutations in bacteria: high rate and small effects. *Science*, 317(5839):813–815, August 2007.

[45] A.D. Peters and S.P. Otto. Liberating genetic variance through sex. *BioEssays*, 25:533–537, 2003.

[46] E. Pollack. Fixation probabilities when the population size undergoes cyclic fluctuations. *Theoretical Population Biology*, 57:51–58, 2000.

[47] Sean H. Rice. Evolutionary theory: mathematical and conceptual foundations. *Sinauer Associates Inc Publishers*, 1961.

[48] 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. *J. Mol. Evol.*, 67:368–376, 2008.

[49] 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. *Nat. Genet.*, 37:441–444, 2005.

[50] I.M. Rouzine, E. Brunet, and C.O. Wilke. The travelling-wave approach to asexual evolution: Muller's ratchet and the speed of adaptation. *Theoretical Population Biology*, 73:24–46, 2008.

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

[52] L.M. Wahl and P.J. Gerrish. The probability that beneficial mutations are lost in populations with periodic bottlenecks. *Evolution*, 55(12):2606–2610, December 2001.

[53] C.O. Wilke. The speed of adaptation in large asexual populations. *Genetics*, 167:2045–2053, 2004.