Population Dynamics in the Presence of Quasispecies Effects and Changing Environments

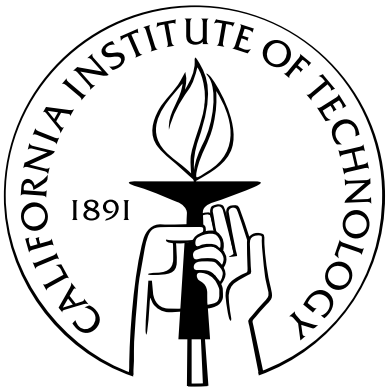
Thesis by

Robert Forster

In Partial Fulfillment of the Requirements

for the Degree of

Doctor of Philosophy



California Institute of Technology

Pasadena, California

2006

(Defended 21 April 2006)

Acknowledgments

First and foremost, I would like to thank my advisor, Chris Adami, for creating the challenging, inspirational, and open environment that was the Digital Life Lab at Caltech. It has been an amazing place to work and this thesis would not have been possible otherwise.

I would also like to thank Claus Wilke, my mentor and collaborator, for his encouragement, patience, and most of all for sharing his knowledge and clear understanding of evolutionary dynamics. In addition, I would like to thank my committee, Emlyn Hughes, Niles Pierce, and Mark Wise, for their interesting and constructive comments. The other members of our group, Jesse Bloom, Stephanie Chow, Evan Dorn, Allan Drummond, and Alan Hampton, were all helpful in numerous discussions, mathematical, computational, and otherwise. Brian Baer at Michigan State has my gratitude for all his help these past years running my various simulations on his supercomputer.

On a personal note, I wanted to thank my many friends at Caltech for their support, friendship, hospitality, and even occasional insights into my research. This includes among others Sumit Daftuar, Keith Matthews, Matt Matuszewski, and Shanti Rao. A big thank you from the bottom of my stomach to all members past and present of the Prufrock Dinner Group for sharing their excellent cooking and tolerating mine. My parents, Lynn and Bob, and my sister Cara deserve particular thanks for their love, encouragement, and high expectations throughout the years. Lastly, my wife Chi has my thanks for all her love, support, and assistance, most especially while I finished writing this thesis.

Abstract

This thesis explores how natural selection acts on organisms such as viruses that have either highly error-prone reproduction or face variable environmental conditions or both. By modeling population dynamics under these conditions, we gain a better understanding of the selective forces at work, both in our simulations and hopefully also in real organisms. With an understanding of the important factors in natural selection we can forecast not only the immediate fate of an existing population but also in what directions such a population might evolve in the future.

We demonstrate that the concept of a quasispecies is relevant to evolution in a neutral fitness landscape. Motivated by RNA viruses such as HIV, we use RNA secondary structure as our model system and find that quasispecies effects arise both rapidly and in realistically small populations. We discover that the evolutionary effects of neutral drift, punctuated equilibrium and the selection for mutational robustness extend to the concept of a quasispecies. In our study of periodic environments, we consider the tradeoffs faced by quasispecies in adapting to environmental change. We develop an analytical model to predict whether evolution favors short-term or long-term adaptation and validate our model through simulation. Our results bear directly on the population dynamics of viruses such as West Nile that alternate between two host species. More generally, we discover that a selective pressure exists under these conditions to fuse or split genes with complementary environmental functions. Lastly, we study the general effects of frequency-dependent selection on two strains competing in a periodic environment. Under very general assumptions, we prove that stable coexistence rather than extinction is the likely outcome. The population dynamics of this system may be as simple as stable equilibrium or as complex as deterministic chaos.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Evolution and Population Dynamics

Ever since Darwin's theory of "survival of the fittest" (3), biologists and ecologists have sought to understand and predict the variations in natural populations. The early 20th century saw the application of mathematical models to the field of population dynamics, with such figures as Wright, Fisher, and Lotka among others making significant contributions (6, 20, 13). Later Kimura's application of the mathematics of partial differential equations to population dynamics would lead to considerable advances in understanding (11). Kimura's theory of neutral evolution emphasized the importance that superficially insignificant, or neutral, changes could have on a population's global evolution (12). While not widely appreciated at the time, the importance of neutral changes in the process of adaptation would be a recurring theme in later studies (9, 10).

With the arrival of cheap and increasingly powerful computers, the study of population dynamics is no longer restricted to analytic mathematical modeling applied to a handful of tractable special cases. The ability to perform quick and numerous simulations has facilitated the study of increasingly complex situations. Real world complexities such as competition between many species, complicated interspecies interactions, or evolution under variable environmental conditions have become accessible topics of study. While analytic solutions to such complex situations remain rare, heuristic approaches and physically motivated approximations can now be validated with extensive simulation results.

This thesis extends our understanding of population dynamics in cases of varying environments and when numerous neutral changes are possible. Through a combination of analytic work, approximate methods, and numerical simulation, I aim to clarify the important factors that determine the course of evolution under these conditions. The models developed and the insights obtained from these investigations are most directly applicable to the study of viral populations, an area of particular interest in light of the recent spread and risks posed by West Nile virus and avian influenza among others.

What follows is a brief discussion and historical overview of three relevant concepts from population dynamics. This introduction will cover the topics to be investigated in more detail in the subsequent chapters of this thesis.

1.1.1 The Quasispecies Concept

The concept of a *quasispecies* was first formulated by Eigen in 1971 (4). Eigen originally considered chemical species in the situation where reactions could convert between the different types of chemical species at given rates. The quasispecies he described in this context is the equilibrium distribution of chemical species. In the context of population dynamics, the term quasispecies refers to an equilibrium distribution of closely related biological species. A quasispecies can only arise in the situation where the species studied are coupled by potential reproductive mutations. For example, when studying species A and B, it would be necessary that a member of A could accidentally produce an offspring of species B (or vice versa) for the formation of a quasispecies. This seems fairly implausible when applied to higher life forms, where taking A = cat and B = dog would require a cat to give birth to puppies or a dog to kittens! Quasispecies effects are typically relevant on a more microscopic level, where we associate the DNA or RNA genome of an organism as uniquely defining its species or strain.[1] For example, if a virus makes a mistake in its molecular copying machinery and the resulting offspring's genome has an adenine swapped for a thymine, this gives rise to a slightly different genome and hence a different strain. Depending on whether or not this

[1] The concept of a species is relatively well defined for higher organisms in terms of whether two organisms can successfully interbreed. For asexual organisms such as bacteria or viruses, we use *strains* to refer to different types of related individuals and avoid the semantic issues associated with the use of the term "species."

change on the DNA level gives rise to a phenotypic (observable) difference is a separate issue that we will consider shortly.

As originally conceived by Eigen, a quasispecies was associated with a special strain called the "master sequence." While the quasispecies concept has turned out to be much more broadly applicable, this original view can serve as an instructive example. In the case of a master sequence, the reproductive rate of this special strain is high, whereas all other strains are defective and reproduce more slowly. Without any mutational coupling between strains, the master strain would rapidly outcompete all other strains and take over any population. In that case there would be no quasispecies. However, when mutations are possible between strains, the largely successful master sequence population will inevitably produce a few mutants in the course of its rapid and error-prone reproduction. The equilibrium distribution of the master sequence together with its mutants are what we refer to as a quasispecies. If the probability of mutations is small, the influence of these mutants will be minor and the equilibrium population will consist almost entirely of copies of the master sequence. However, if the mutation rate in reproduction is high, the equilibrium distribution of strains may even have more total mutants in the population than members of the master strain. For example, figure 1.1A shows the fitness of a master strain that has a considerable fitness advantage relative to the other strains. The quasispecies that arise at different mutation rates are shown for this example in figure 1.1B,C,D.

More generally, a quasispecies may be present even in the absence of a master sequence. If all sequences are equally fit, we say that all sequences are *neutral*, referring to the lack of any differences in reproductive fitness. In this case, it is useful to think of all the strains arranged in a *neutral network*. This network is a graph in which each strain appears as a vertex and edges connect any two vertices if those strains can produce each other as a mutant during reproduction. A simple example of a neutral network is shown in figure 1.2. Despite the lack of differing reproductive fitness across strains in a neutral network, the quasispecies that forms may still have certain strains represented at higher abundances than others. These differences can arise from inhomogeneities in the structure of possible mutations. For example, this could happen if a certain strain makes

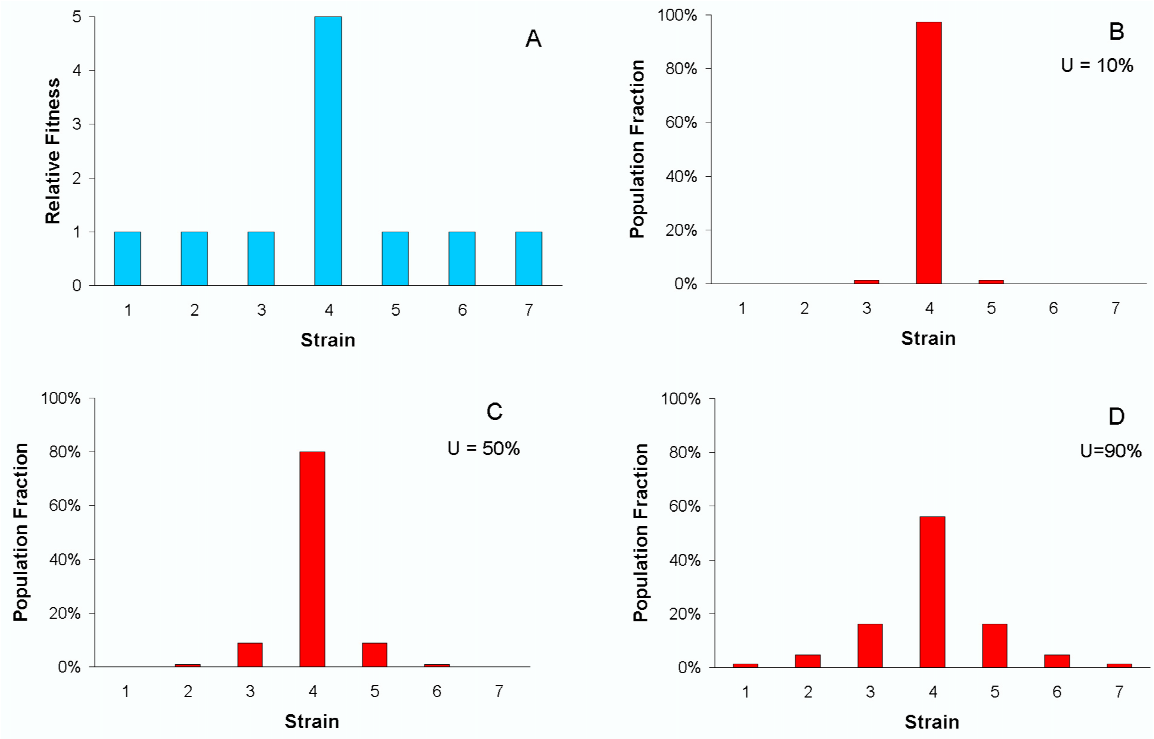


Figure 1.1: A: The relative fitness of each of seven strains, where the center one (strain 4) is the master sequence. B, C, D: Equilibrium population distribution of the strains, given probabilities of reproductive error $U = 10\%$, 50%, or 90% respectively. A reproductive error is equally likely to shift the strain of the offspring by one in either direction.

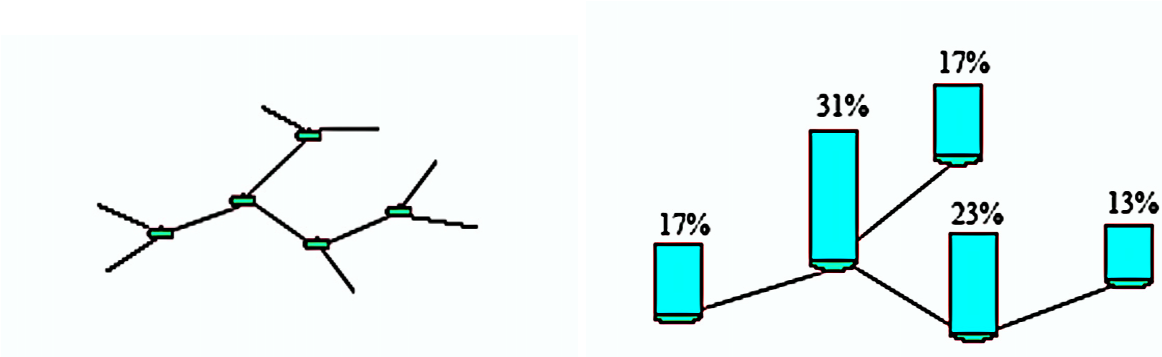


Figure 1.2: Left: a small neutral network of five related strains. Right: The equilibrium population distribution, or quasispecies, associated with this network.

reproductive mistakes less frequently than others strains, or if a certain strain is more likely to arise from a mutation than others. This latter case corresponds to being in a "highly connected" region of the graph, and an example of this is shown in figure 1.2. If $x(t)$ represents a vector containing the population fraction of each of the five nodes in figure 1.2 at a given time t, we find $x(t+1)$ as follows:

$$x(t+1) = \begin{bmatrix} 1-U & U/3 & U/3 & U/3 & 0 \\ U/3 & 1-U & 0 & 0 & U/3 \\ U/3 & 0 & 1-U & 0 & 0 \\ U/3 & 0 & 0 & 1-U & 0 \\ 0 & U/3 & 0 & 0 & 1-U \end{bmatrix} x(t) \tag{1.1}$$

The diagonal entries of the transition matrix correspond to the probability of error-free reproduction, while the off-diagonal entries correspond to the possible mutations allowed in the neutral network. After a long time, the initial population distribution will converge to the eigenvector with the largest eigenvalue. This eigenvector represents the quasispecies distribution and is shown in figure 1.2.

1.1.2 Changing Environments

Variable environmental conditions are the norm among natural populations. Elton's work in the early 20th century highlighted the important effects of climatic variation on numerous animal populations (5). Diurnal effects, the changing seasons, and even global warming are all classic examples of

environmental change. The body's immune response to a viral infection represents a changing environment from the perspective of the virus, but unlike the seasons, this environmental change is a direct consequence of the presence of the virus itself. Although both types of environmental changes are of interest, we are primarily concerned with those of the former type where the actions of the species in question do not result in feedback to the environment.

In a *periodic* environment, the environmental conditions alternate between two different states in a regular pattern. From a mathematical perspective, a periodic environment can be treated in much the same way as a static one. After allowing sufficient time to ignore transient effects, the population still reaches an equilibrium distribution although this distribution is now a function of the phase of the cycle. There is a quite good analogy between the influence of environmental changes on the population and the action of a low-pass filter (19, 18). Just as a low-pass filter averages over high frequencies, an evolving population faced with a rapidly changing environment adapts to the average environmental conditions rather than specifically to the current environment. Likewise, just as low frequencies are unchanged by the filter, environmental changes that happen slowly enough pass through to the population and the population adapts to the current state of the environment. In this analogy, the rate of environmental change is fast or slow relative to the population's ability to adapt.

1.1.3 Frequency-Dependent Selection

Frequency-dependent selection refers to the situation where the competitive advantage one species has over another is not constant but instead varies depending on the relative abundances, or frequencies, of the species in question. Host-parasite interactions are a common example of this (15), where the host's fitness suffers at high levels of parasitism and likewise the parasite's fitness suffers when the host species becomes scarce. Other examples include predator-prey interactions and the selection for differing behavioral traits within a single species, such as aggression level or mating strategies.

With Nicholson's and Haldane's early work shedding light on the importance of frequency-

dependent effects in natural systems (16, 8), this has been an area of considerable interest. Models of frequency dependence have been studied as a mechanism for maintaining genetic variation (2). The population dynamics that arise under conditions of frequency-dependent selection can vary considerably. On one hand, the selective forces could be stabilizing and lead to an equilibrium between the two strains considered. Alternatively, exceedingly complex outcomes are also observed, with periodic oscillations and deterministic chaos being possible (14). Even very simple model systems can lead to highly nonlinear phenomena and chaotic dynamics (1, 7).

1.2 Thesis Overview

The research done for this thesis can be divided into three distinct categories involving the study of population dynamics in the presence of quasispecies effects or changing environments.

1.2.1 Evolution in Finite Quasispecies

Quasispecies are typically defined in terms of the equilibrium population distribution of a group of related strains. This definition in terms of equilibrium requires the physically unreasonable assumption of an infinite population size. In chapter 2 we demonstrate the presence and importance of quasispecies effects in more modestly sized populations as small as 30–100 individuals. Our model system for this study is RNA sequences where the minimum free energy shape of each RNA sequence, specifically its secondary structure, is used to determine a sequence's fitness. In studying how evolution acts in this system, we observe that not only are quasispecies present at small population sizes but also that their evolution shows selection for beneficial mutations and neutral drift, features commonly associated with natural selection acting in a more traditional context. Our results suggest that the quasispecies concept is applicable to the population dynamics of many RNA viruses.

1.2.2 Finite Quasispecies in a Changing Environment

Changing environments are often of practical interest in the study of population dynamics. Although widely studied in other contexts, variable environments have only recently been applied to

quasispecies and only in the context of infinite populations (17, 19, 18). In chapter 3 and chapter 4, we study how natural selection acts on two finite quasispecies competing in a periodic environment. We explore the tradeoffs these quasispecies face between short-term and long-term adaptation to the environmental conditions and develop a model to predict the best adaptive strategy. Relative to the infinite quasispecies case, we discover that qualitatively different and complex evolutionary dynamics arise due to finite quasispecies effects.

1.2.3 Frequency-Dependent Selection in a Changing Environment

Changing environments and frequency-dependent selection are relevant in many natural populations. While each topic separately has been the subject of numerous mathematical models, they are rarely studied together. In chapter 5, we investigate the competition between two specialist strains, in which each strain is well adapted to only one of the two periodic environmental states. Under the simplifying assumption that each strain competes best when rare, we derive general analytic results for the possible outcomes of the competition between the two strains. Coexistence rather than extinction is found to be the likely outcome for a wide range of conditions. The population dynamics describing this coexistence may be simple, although periodic or chaotic oscillations are also possible.

Bibliography

[1] Altenberg, L. (1991). Chaos from linear frequency-dependent selection. *Am. Nat.*, *138*, 51–68.

[2] Cockerham, C. C., Burrows, P. M., Young, S. S., & Prout, T. (1972). Frequency-dependent selection in randomly mating populations. *Am. Nat.*, *106*, 493–515.

[3] Darwin, C. (1859). *On the Origin of Species by Means of Natural Selection*. London: John Murray.

[4] Eigen, M. (1971). Self-organization of matter and the evolution of macromolecules. *Naturwissenschaften*, *58*, 465–523.

[5] Elton, C. S. (1924). Periodic fluctuations in the numbers of animals: Their causes and effects. *J. Exp. Biol.*, *2*, 119–163.

[6] Fisher, R. A. (1930). *The Genetical Theory of Natural Selection*. Oxford: Oxford University Press.

[7] Gavrilets, S., & Hastings, A. (1995). Intermittency and transient chaos from simple frequency-dependent selection. *Proc. R. Soc. Lond. B*, *261*, 233–238.

[8] Haldane, J. B. S. (1953). Animal populations and their regulation. *New Biology*, *15*, 9–24.

[9] Huynen, M. A. (1996). Exploring phenotype space through neutral evolution. *J. Mol. Evol.*, *43*, 165–169.

[10] Huynen, M. A., Stadler, P. F., & Fontana, W. (1996). Smoothness within ruggedness: The role of neutrality in adaptation. *Proc. Natl. Acad. Sci. USA*, *93*, 397–401.

[11] Kimura, M. (1964). Diffusion models in population genetics. *J. Appl. Prob.*, *1*, 177–232.

[12] Kimura, M. (1983). *The neutral theory of molecular evolution*. Cambridge: Cambridge University Press.

[13] Lotka, A. J. (1925). *Elements of physical biology*. Baltimore: Williams and Wilkins.

[14] May, R. M. (1979). Bifurcations and dynamic complexity in ecological systems. *Ann. N. Y. Acad. Sci.*, *316*, 517–529.

[15] May, R. M., & Anderson, R. M. (1983). Epidemiology and genetics in the coevolution of parasites and hosts. *Proc. R. Soc. Lond. B*, *219*, 281–313.

[16] Nicholson, A. J. (1954). An outline of the dynamics of animal populations. *Aust. J. Zool.*, *2*, 9–65.

[17] Nilsson, M., & Snoad, N. (2000). Error thresholds on dynamic fitness landscapes. *Phys. Rev. Lett.*, *84*, 191–194.

[18] Nilsson, M., & Snoad, N. (2002). Quasispecies evolution on a fitness landscape with a fluctuating peak. *Phys. Rev. E*, *65*, 031901.

[19] Wilke, C. O., Ronnewinkel, C., & Martinetz, T. (2001). Dynamic fitness landscapes in molecular evolution. *Phys. Rep.*, *349*, 395–446.

[20] Wright, S. (1931). Evolution in Mendelian populations. *Genetics*, *16*, 97–159.

Chapter 2

Quasispecies Can Exist Under Neutral Drift at Finite Population Sizes

Submitted to *Journal of Theoretical Biology*, August 2005.

Authors as published: Robert Forster, Christoph Adami, and Claus O. Wilke.

2.1 Abstract

We investigate the evolutionary dynamics of a finite population of RNA sequences adapting to a neutral fitness landscape. Despite the lack of differential fitness between viable sequences, we observe typical properties of adaptive evolution, such as increase of mean fitness over time and punctuated-equilibrium transitions. We discuss the implications of these results for understanding evolution at high mutation rates, and extend the relevance of the quasispecies concept to finite populations and time scales. Our results imply that the quasispecies concept and neutral drift are not complementary concepts, and that the relative importance of each is determined by the product of population size and mutation rate.

2.2 Introduction

The quasispecies model of molecular evolution (14, 16) predicts that selection acts on clouds of mutants, the *quasispecies*, rather than on individual sequences, if the mutation rate is sufficiently high. RNA viruses tend to have fairly high mutation rates (12, 13), and therefore the quasispecies model is frequently used to describe the evolutionary dynamics of RNA virus populations (7, 10, 9, 8). However, this use has generated criticism (26, 31), because quasispecies theory, as it was originally developed, assumes an infinite population size and predicts deterministic dynamics. Viral populations, on the other hand, are finite and subject to stochastic dynamics and neutral drift.

However, the hallmark of quasispecies dynamics—the existence of a mutationally coupled population that is the target of selection in its entirety—does not presuppose an infinite population size or the absence of neutral drift (5, 36, 44, 50). Rather, infinite populations were used by Eigen (14) and Eigen and Schuster (16) to simplify the mathematics of the equations describing the population dynamics. Even though technically, the quasispecies solution of Eigen and Schuster, defined as the largest eigenvector of a suitable matrix of transition probabilities, exists only for infinite populations after an infinitely long equilibration period, it would be wrong to conclude that the cooperative population structure induced by mutational coupling disappears when the population is finite. We

show here that quasispecies dynamics are evident in fairly small populations (effective population size $N_e \leq 1000$), and that these dynamics cross over to pure neutral drift in a continuous manner as the population size decreases.

We simulate finite populations of self-replicating RNA sequences and look for an unequivocal marker for quasispecies dynamics in this system, the selection of mutational robustness (44, 2, 48, 53). We choose RNA secondary structure folding (25) as a fitness determinant because it is a well-understood model in which the mapping from sequence to phenotype is not trivial. The nontriviality of this mapping is crucial for the formation of a quasispecies, as we will explain in more detail later.

Since the existing literature on the evolution of RNA secondary structures is extensive, we will now briefly review previous works and then describe how our study differs. We can subdivide the existing literature broadly into three categories: (i) studies that investigate how secondary structures are distributed in sequence space; (ii) studies that investigate how sequences can evolve from one structure into another; and (iii) studies that investigate the evolution of sequences that all fold into the same secondary structure, that is, the evolution of sequences on a single neutral network.

Studies in the first category have established the importance of neutral networks for RNA secondary structure folding (40, 39). All sequences folding into the same secondary structure form a network in genotype space, that is, a graph that results from including all these sequences as vertices, and including an edge between two such vertices if a single mutation can interconvert the two sequences. The number of edges connected to a vertex is called the degree of neutrality of that vertex (i.e., sequence). These neutral networks span large areas of sequence space, the neutral networks' size distribution follows a power law, and for any two secondary structures of comparable size, there are areas in sequence space in which sequences folding into both structures can be found in close proximity (40, 21, 22, 39). Further, the fitness landscapes derived from RNA secondary structure folding are similar to basic models of fitness landscapes, but differ in details (19, 4), and are highly epistatic (52, 54).

The main result from studies investigating the evolution of secondary structures is that evolution proceeds in a stepwise fashion: A single secondary structure dominates the population for an

extended period of time (an epoch), but intermittently a new, improved structure will appear and take over the population (30, 17, 18). During the epochs when the population is seemingly static, the population diffuses over the neutral network of the currently dominant structure. It is primarily because of this prolonged diffusion that the population has a chance to discover a new structure with higher fitness (28, 30, 17, 18). Details of the diffusion process and the transition probabilities from one structure to another have been worked out (20, 30, see also next paragraph).

We can interpret studies in the third category as describing the evolutionary dynamics during the epochs of phenotypic stasis observed in the evolution of secondary structures. As already mentioned, the sequences diffuse over the neutral network, and any specific sequence is rapidly lost from the population (30, 38). However, the sequences do not diffuse as a single, coherent unit, but instead form separate clusters that diffuse independently of each other (20, 30). It is useful to extend Eigen's concept of the error threshold (14) to distinguish between the genotypic error threshold, that is, the mutation rate at which a specific sequence cannot be maintained in the population, and the phenotypic error threshold, that is, the mutation rate at which a given secondary structure cannot be maintained in the population (20, 38). For RNA, the genotypic error threshold occurs usually for an infinitesimally small positive mutation rate, while the phenotypic error threshold occurs at fairly large mutation rates (20, 38). The exact position of the phenotypic error threshold depends on the size of the neutral network and the fitness of suboptimal secondary structures (38).

It is important to distinguish the diffusion over a neutral network from drift in a completely neutral fitness landscape (6, 23, 24). If the product of population size and mutation rate is sufficiently high, then on a neutral network (where a fraction of all possible mutations is deleterious) there is a selective pressure that keeps the population away from the fringes of the neutral network, and pushes it towards the more densely connected areas in the center of the neutral network (44, 2, 48). This selective pressure has been termed "selection for mutational robustness" (44), and is a tell-tale sign that selection occurs in the quasispecies mode on clouds of mutants, rather than on individual sequences. Van Nimwegen et al. and Bornberg-Bauer and Chan were the first to develop a formal theory for this effect, but anecdotal evidence for it had been observed previously (29, 20). The

theory developed by van Nimwegen et al. and Bornberg-Bauer and Chan applies only to infinite populations. Nevertheless, simulations have shown that this effect occurs also in large but finite populations if the mutation rate is sufficiently large (44, 50).

According to the quasispecies model, mutational robustness is as important a component of fitness as is replication speed (41, 55, 49). This observation suggests that a sudden transition to increased mean fitness may not only be caused by the discovery of a sequence with higher replication rate, but also by the discovery of a more densely connected region of the neutral network the population is already residing on, without any obvious change in the sequences' phenotype (48). Here, we study these types of transitions—which change the population mean fitness while the secondary structure remains unchanged—as they represent the ultimate demonstration of quasispecies selection. In our simulations, we consider all RNA sequences that fold into a specific target secondary structure as viable. All viable sequences have the same fitness, arbitrarily set to one. All RNA sequences that do not fold into the target secondary structure are nonviable, with fitness 0. There is no phenotypic error threshold in our simulations, and the target structure can only be lost from the population through sampling noise. The latter outcome is extremely unlikely for all but the smallest population sizes. However, if it occurs the population dies, because all remaining sequences have fitness zero. Our choice of fitness landscape guarantees that all changes that we see in mean fitness must be caused by changes in the population neutrality. (Here and in the following, we refer to the average degree of neutrality among viable members of the population as the population neutrality or simply the *neutrality*.)

2.3 Materials and Methods

We consider a population of fixed size N composed of asexual replicators whose probability of reproduction in each generation is proportional to their fitness (Wright-Fisher sampling). The members of the population are RNA sequences of length $L = 75$, and their fitness w is solely a function of their secondary structure. Those that fold into a specific target secondary structure (such as figure 2.1) are deemed viable with fitness $w = 1$, while those that fold into any other

shape are nonviable ($w = 0$). The average fitness $\langle w \rangle$ of the population is therefore the fraction of living members out of the total population. RNA sequences are folded into the minimum free energy structure using the Vienna Package (25), and dangling ends are given zero free energy (46). For a given simulation, an initial RNA sequence is selected uniformly at random and its minimum-energy secondary structure defines the target structure for this simulation, thereby determining a neutral network on which the population evolves for a time of $T = 50{,}000$ generations. Mutations occur during reproduction with a fixed probability μ per site, corresponding to an average genomic mutation rate $U = \mu L$.

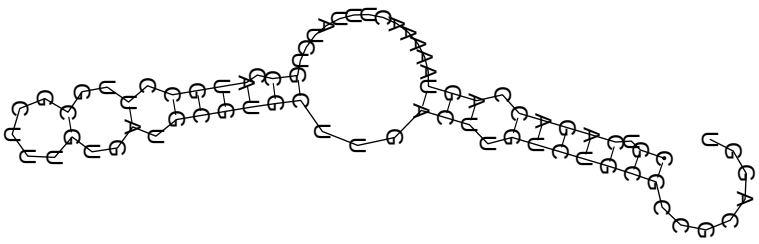


Figure 2.1: The minimum free energy secondary structure of the RNA sequence shown.

Our simulations spanned a range of genomic mutation rates and population sizes, and we performed 50 independent replicates for each of the pairs (U, N), starting each with a different randomly chosen initial sequence. To study mutation rate effects, we considered a fixed population size of $N = 1000$, across a range of genomic mutation rates, using $U = 0.1, 0.3, 0.5, 1.0$, and 3.0. To study effects due to finite population size, we considered a fixed mutation rate of $U = 1.0$, using population

sizes of $N = 30, 100, 300$, and 1000.

The degree of neutrality of a sequence was determined by calculating the fraction of mutations that did not change the minimum-energy secondary structure. Thus, if N_ν of all $3L$ one-point mutants of a sequence retain their structure, the degree of neutrality of that sequence is given by $\nu = N_\nu/3L$. Because sequences that do not fold into the target structure have zero fitness, a sequence's degree of neutrality is equal to the mean fitness of all possible single mutants. We recorded the population's average fitness every generation, while the population's average neutrality, being much more computationally expensive, was calculated only at the start and end of each replicate. For illustrative purposes, select replicates of interest were recreated using the original random seed, and the population neutrality was recorded every 100 generations.

To observe the signature of natural selection acting within our system, we derived a statistical approach to identify transitions in the population's average fitness $\langle w \rangle$. If a beneficial mutation appears and is subsequently fixated in the population, we expect to observe a step increase in the population's average fitness. We emphasize again that such selective sweeps must be due to periodic selection of quasispecies for increased mutational robustness, since there are no fitness differences between individual genotypes.

In light of the fluctuations in the population's average fitness due to mutations and finite population effects, we employed statistical methods to estimate the time at which the increase in average fitness occurred and associated a p-value with our level of confidence that a transition has occurred. Our approach can be thought of as a generalization of the test for differing means between two populations (those before and after the mutation), except that the time of the mutation's occurrence is unknown *a priori*. For a full derivation and discussion of our approach, see appendix A. While our algorithm can be applied recursively to test for and identify multiple transitions that may occur in a single simulation, unless otherwise noted, we considered only the single most significant transition found.

2.4 Results

Because replicates were initialized with N (possibly mutated) offspring of the randomly chosen ancestor, the simulation runs did not start in mutation–selection balance. Typically, we observed an initial equilibration period of 50 to 200 generations, after which the population's fitness and neutrality stabilized, with fluctuations continuing with magnitude in proportion to the mutation rate. As predicted by van Nimwegen et al. (44), during the equilibration period we observed in most replicates beneficial mutations that increased the equilibrium level of both average fitness and neutrality. (Throughout this paper, by beneficial mutations we mean mutations that increase a sequence's degree of neutrality, and thus indirectly the mean fitness of the population. There are no mutations that increase the fitness of a viable sequence beyond the value 1 in our system.) These mutations led to the initial formation of a quasispecies in a central region of the neutral network. For the remainder of this paper, we are not interested in this initial equilibration, but in transitions towards even more densely connected areas of the neutral network once the initial equilibration has occurred.

To determine if such a transition has occurred, we need a method to distinguish significant changes in the population's mean fitness from apparent transitions caused by statistical fluctuations. We devised a statistical test (see appendix A for details) that can identify such transitions and assign a p-value to each event. We found that transitions to higher average fitness occurred in over 80% of simulations across all mutation rates studied, if we considered all transitions with p-values of $p < 0.05$. Figure 2.2 shows a particularly striking example of such a transition (p-value $\leq 10^{-7}$), where a 5.0% increase in average fitness occurs at $t = 9814$. A similar analysis of the average population neutrality (not usually available, but computed every generation specifically in this case) finds an increase of 11.2% occurring at $t = 9876$, with the same level of confidence. The multiple transitions shown in the figure 2.2 are the results of recursively applying our step-finding algorithm until no steps are found with $p < 0.05$.

Depending on the mutation rate, a step size as little as 0.04% in the population's average fitness could be statistically resolved in a background of fitness fluctuations several times this size. For

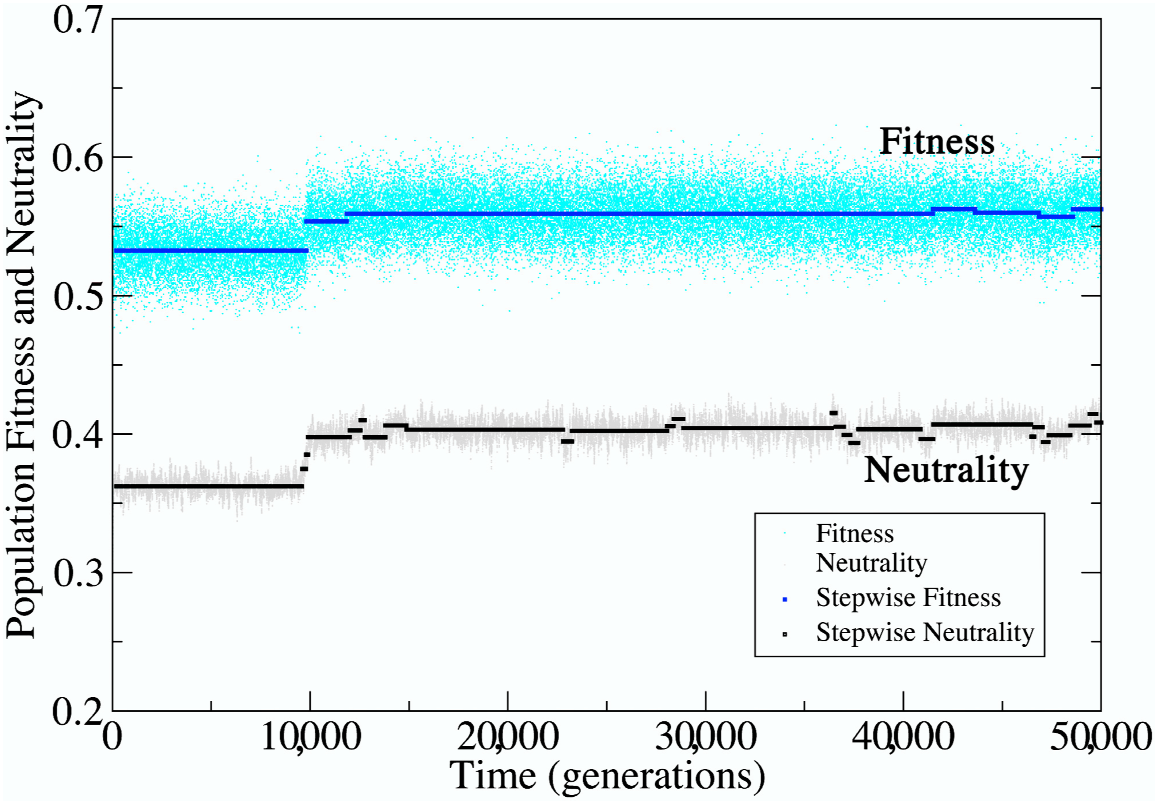


Figure 2.2: Average fitness and neutrality of a population during a single simulation at a genomic mutation rate of $U = 1.0$. All live members of the population fold into the secondary structure shown in figure 2.1. At $t = 9814$, a 5% increase in the population's average fitness occurs at the $p < 10^{-7}$ level, with a corresponding transition in the population's average neutrality. Smaller transitions occur throughout the simulation run. The solid lines indicate the epochs of constant fitness and neutrality, as determined by our step-finding algorithm. As explained in appendix A, the application of this algorithm to the neutrality data is for illustrative purposes only. Because of temporal autocorrelations in the neutrality, not all steps that the algorithm identifies are statistically significant.

comparison, typical noise levels, as indicated by the ratio of the standard deviation of the fitness to its mean, ranged from 0.7% to 6.6% over the mutation rates studied. Note that fluctuations in the population's neutrality level are much smaller, due to the additional averaging involved. However, because neutrality is much more expensive computationally, and would also be difficult to measure in experimental viral populations, we used mean fitness as an indicator of transitions throughout this paper.

Figure 2.3 shows the average size of the most significant step observed as a function of the mutation rate. At low mutation rates, such as $U = 0.1$, the smaller observed step size corresponds to the fact that 90% of the population is reproducing without error, and hence improvements in

neutrality can only increase the population's fitness in the small fraction of cases when a mutation occurs. At higher mutation rates the step sizes increase, reflecting the larger beneficial effect of increased neutrality under these conditions.

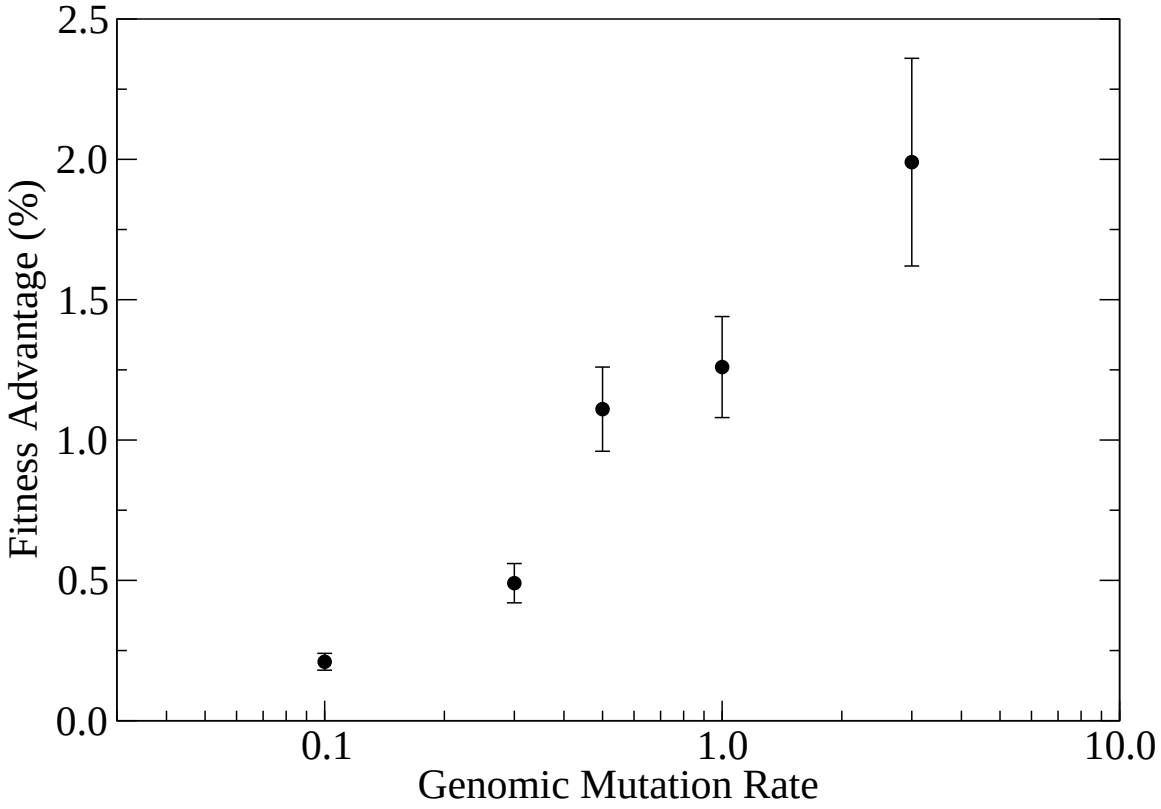


Figure 2.3: Average step size as a function of genomic mutation rate ($U = 0.1, 0.3, 0.5, 1.0, 3.0$). Step size is measured by percent increase in the population's fitness, with only runs significant at the $p < 0.05$ level shown. Error bars are standard error.

In about 10% of all simulations with statistically significant changes in fitness, the most significant change in fitness was actually a step *down*, that is, a fitness loss, rather than the increase in fitness typically observed. Negative steps in average fitness occur due to stochastic fixation of detrimental mutations at small population sizes (33). These negative fitness steps, however, are generally much smaller than the typical positive step size. The average size of these negative steps was between 0.09% and 0.77%, compared with an average positive step size between 0.27% and 2.33% (see figure 2.4).

We specifically studied the role of finite population size and its effects on neutral drift by considering populations of size $N = 30$, 100, 300, and 1000 at a constant genomic mutation rate of $U = 1.0$. We again performed 50 replicates at each population size, and the distribution of statisti-

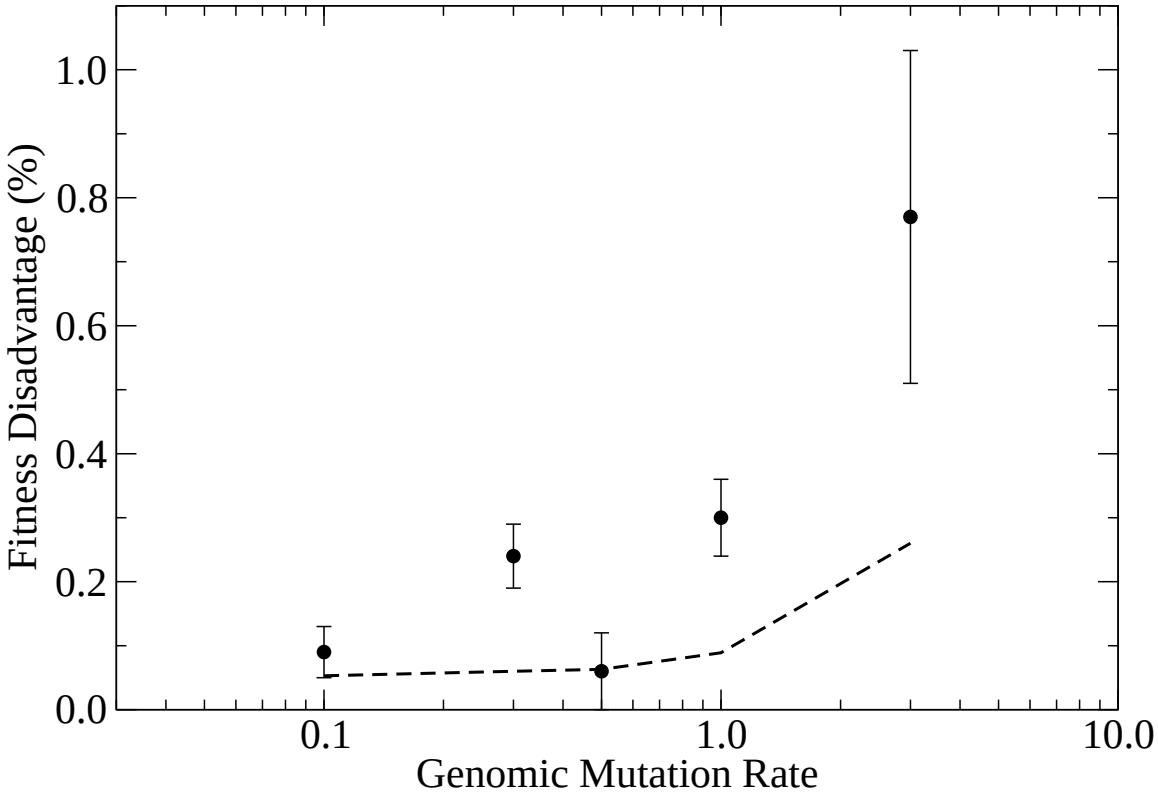


Figure 2.4: Average step size $|s|$ of statistically significant drops in fitness (at the $p < 0.05$ level). Step size is measured by relative decrease in population fitness, and error bars are standard error. The dotted line indicates $2|s| = 1/N_e$, a selective disadvantage consistent with neutral drift in a finite population. N_e is the average number of living members of the population (effective population size).

cally significant step sizes are shown in fig. 2.6 (biggest step only) and fig. 2.7 (all steps). While the larger population's distributions show a clear bias towards positive steps in fitness, the distributions become increasingly symmetric about zero for smaller population sizes. A gap around zero fitness change becomes increasingly pronounced in smaller populations, as the fluctuations in fitness due to finite population size preclude us from statistically distinguishing small step sizes from the null hypothesis that no step has occurred.

We also kept track of the consensus sequence in our simulations, to determine whether the population underwent drift while under selection for mutational robustness. In the runs with $N = 1000$, the consensus sequence accumulated on average one substitution every 2 to 3 generations. As such rapid change might be caused by sampling effects, we also studied the speed at which the consensus sequence changed over larger time windows. Using this method with window lengths of 50

and 100 generations, we found that the consensus sequence accumulated one substitution every 10 to 20 generations (window size 50 generations) or 15 to 30 generations (window size 100 generations). Thus we find that the populations continue to drift rapidly throughout the simulation runs, and never settle down to a stable consensus sequence. Figure 2.5 shows the evolution of the consensus sequence over time for the same simulation run as shown in fig. 2.2.

Time	Consensus Sequence
0	CGUCAGACCAGUAAAAACUUUAUCUGCCAUGCCUUGCGCUUUGUGAUGCGUGCUUGACUUGUCUGCGCCGCAGGU
200	--------A--------AAAA-AA-------------AA-AA----A--A----AC---A-------A-AA-AAA
1000	--A----------U--------C--------A------A--------------A-----------A------
2000	--------G-----A---U----A--
3000	--------C---------A----C--
4000	--------A--------------A--------C----------------------C---C-------U-------
5000	----------------U--------------A----------------------A---A-------A-----C-
6000	------U-------U-A--------------C----------------------G----A-----C-C---A-
7000	--------------AC-----------------------C--------------A----------A-A------
8000	---------------A----------------A----U--A------------------------C-------
9000	----------------------------A--C----A-----------U-----------------A-------
10000	------------------C-----A--AU---A-----------------A--A------------U------C
11000	---------U-A------A-----------------------------------U-A--------A------A
12000	----------U---A------------------
13000	--------------------------------AA-------G--G----------------------------
14000	--------------------------------G-------------------------------------U--
15000	--------------------------------A-U-----------------------------------A--
16000	--------------------------------G-A-----------------------------------U---
17000	--A---
18000	--------C---G---G-U-
19000	--------A---A---A-A-
20000	--

Figure 2.5: Change in the consensus sequence over time, from the same simulation run as presented in fig. 2.2. Dots in the alignment indicate that the base at this position is unchanged from the previous line.

Finally, to confirm that our finite population was not sampling the entire neutral network during our simulations, we estimated the average size of the neutral network. We can represent each RNA secondary structure in dot-and-parenthesis notation, where matched parentheses indicate a bond between the bases at those points in the sequence and dots represent unpaired bases. The number of valid strings of length L can be counted using Catalan numbers $\mathrm{Cat}(n) = \binom{2n}{n}/(n+1)$, which give the number of ways to open and close n pairs of parentheses (43). Since there are 4^L possible RNA sequences, we obtain for the average network

$$\langle \text{network size} \rangle = 4^L \Big/ \sum_{i=0}^{[L/2]} \mathrm{Cat}(i) \binom{L}{L-2i} \approx 1.1 \times 10^{12} \tag{2.1}$$

for $L = 75$. While there are known to be about 1.8^L structures for large L (40), eq. (2.1) gives a much better bound for our relatively short sequence length. Furthermore, the above expression is a lower bound to the true average network size, because the denominator counts some unphysical structures, such as hairpins with fewer than 3 bases. For comparison, the number of possible distinct genotypes that can appear in each simulation is maximally $NT = 5 \times 10^7$.

2.5 Discussion

In the study of varying mutation rates, the observed increases in the population's fitness in almost all replicates demonstrate the action of natural selection. Since all viable sequences are neutral and hence enjoy no reproductive fitness advantage, this selection acts on increasing the population's robustness to mutations through increases in its average neutrality (as seen in figure 2.2). Thus, these results show evidence that a quasispecies is present in almost all cases, even though the difference between a randomly drifting swarm and a population structured as a quasispecies decreases as the population size and mutation rate decrease. Our results also show evidence of drift leading to the fixation of detrimental mutations in some populations. The negative steps observed (figure 2.4) were comparable in size to $1/N_e$, the probability of a neutral mutation drifting to fixation.

In the study of varying population sizes, the distribution of mutational effects on fitness showed

an increasing bias towards beneficial rather than detrimental mutations as the population's size increased (figures 2.6, 2.7). At population sizes 100, 300, and 1000, the clear positive bias of mutational effects illustrates the presence of a quasispecies, where natural selection is able to act to improve the population's neutrality and hence its robustness to mutations. As the fluctuations in fitness due to small population size become more significant, selection for neutrality becomes less relevant when the $1/N_e$ sampling noise exceeds the typical step size of 1%. At the smallest population size of 30, there still seems to be a bias towards beneficial mutations, but the evidence is less clear and more replicates are probably necessary to observe a clear signal of quasispecies dynamics.

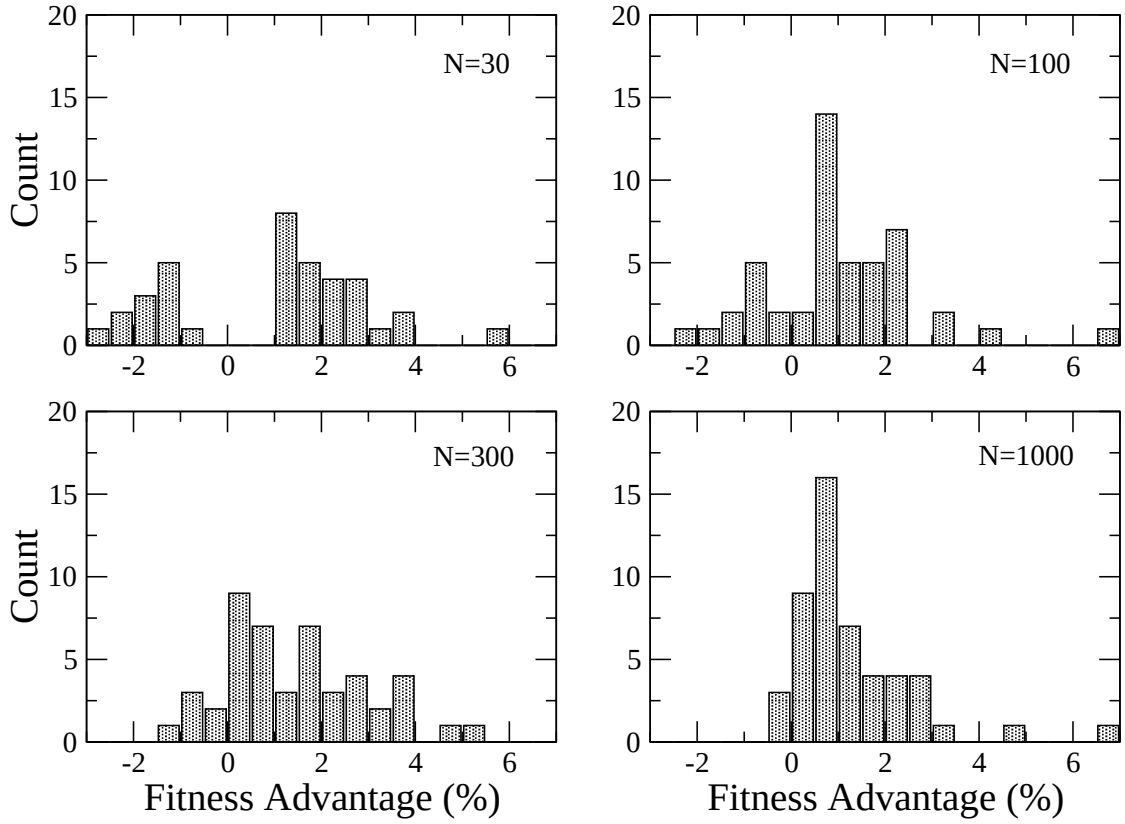


Figure 2.6: Distribution of sizes of the most significant step (at $p < 0.05$) in each run, out of 50 runs at four population sizes ($U = 1$). At small population sizes, the distribution is almost symmetric about zero since most mutations are of less benefit than the $1/N_e$ probability of fixation due to drift. At large sizes, selection is evident from the positively skewed distribution.

Since the average network size is many orders of magnitude larger than the number of sequences

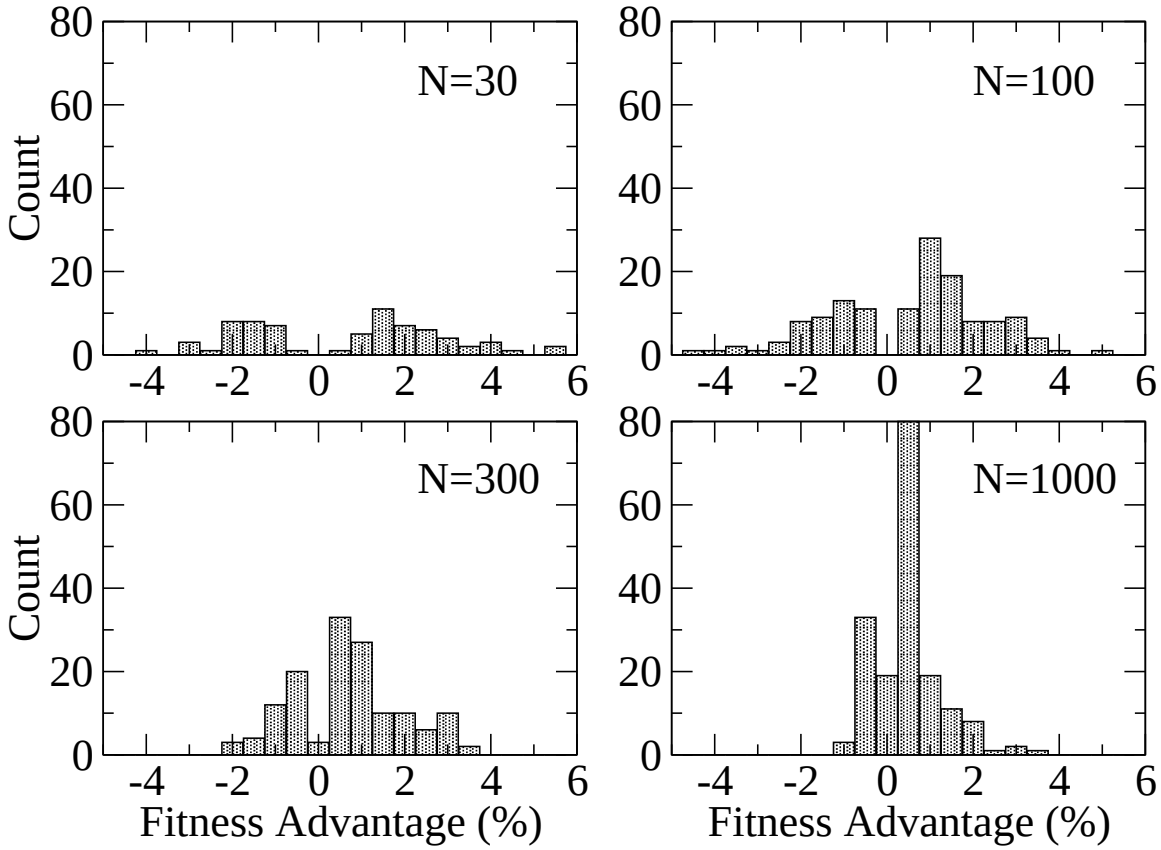


Figure 2.7: Distribution of sizes of all significant steps (at $p < 0.05$) in each run, out of 50 runs at four population sizes ($U = 1$). While these distributions are more symmetrical than those of fig. 2.6, a substantial skew towards positive step sizes is still evident for the larger population sizes.

produced during a simulation, we know that the system is nonergodic and the population cannot possibly have explored the whole neutral network. Moreover, Reidys et al. (39) studied the distribution of neutral network sizes in RNA secondary structure and found that they obey a power law distribution, implying that there are a small number of very large networks, and many smaller networks. As a consequence, choosing an arbitrary initial sequence will more likely result in the choice of a large network. Therefore, eq. (2.1) is effectively a lower bound on the sizes of the networks we actually sampled.

We have shown that quasispecies dynamics is not confined to the infinite population-size limit. Instead, one of the hallmarks of quasispecies evolution—the periodic selection of more mutationally robust quasispecies in a neutral fitness landscape—occurs at population sizes very significantly smaller than the size of the neutral network they inhabit. Despite small population sizes, if the

mutation rate is sufficiently high (in the simulations reported here, it appears that $NU \gtrsim 30$ is sufficient), stable frequency distributions significantly different from random develop on the partially occupied network in response to mutational pressure. Most importantly, we have shown that genetic drift can occur simultaneously with quasispecies selection, and becomes dominant as NU decreases. Thus, the notion that genetic drift and quasispecies dynamics are mutually exclusive cannot be maintained. Instead, we find that both quasispecies dynamics and neutral drift occur at all finite population sizes and mutation rates, but that their relative importance changes.

The existence of a stable consensus sequence in the presence of high sequence heterogeneity has long been used as an indicator of quasispecies dynamics (11, 42, 15, 31, 8). In contrast, the genotypic error threshold for evolution of RNA sequences typically occurs at any small positive mutation rate (20, 30, 38). Here we have shown that quasispecies dynamics can be present while the consensus sequence changes over time. In our simulations, the consensus sequence drifts randomly, in a manner uncorrelated with the transitions in average fitness that we detect. Thus, quasispecies dynamics does not require individual mutants to be stably represented in the population, nor does it require a stable consensus sequence.

The population structure on the neutral network is strongly influenced by the mutational coupling of the genotypes that constitute the quasispecies. This coupling arises because mutations are not independent in the landscape we studied. Rather, as in most complex fitness landscapes, single mutations at one locus can affect the fitness effect of mutations at another (a sign of epistasis, (56)). In the neutral fitness landscape investigated here, mutations at neutral or nonneutral (i.e., lethal) sites can influence the degree of neutrality of the sequence. The absence of epistatic interactions between the neutral mutations in the fitness landscape studied by Jenkins et al. (31) implies the absence of quasispecies dynamics in these simulations. Theoretical arguments show that a non-interacting neutral region in a genome does not alter the eigenvectors of the matrix of transition probabilities, and therefore cannot affect quasispecies dynamics.

Using fitness transitions in neutral fitness landscapes as a tool to diagnose the presence of a quasispecies has a number of interesting consequences from a methodological point of view. Clearly,

because selection for robustness is a sufficient criterion for quasispecies dynamics but not a necessary one, the absence of a transition does not imply the absence of a quasispecies. At the same time, as the population size decreases, fluctuations in fitness become more pronounced, rendering the detection of a transition more and more difficult. Theoretical and numerical arguments suggest that small populations at high mutation rate cannot maintain a quasispecies (44, 48), so the disappearance of the mutational robustness signal at small population sizes is consistent with the disappearance of the quasispecies. However, the type of analysis carried out in this work does not lend itself to detecting quasispecies in real evolving RNA populations, because the fitness landscape there cannot be expected to be strictly neutral. Instead, transitions from one peak to another of different height (3, 35) are likely to dominate. Quasispecies selection transitions such as the one depicted in fig. 2.2 can, in principle, be distinguished from peak-shift transitions in that every sequence before and after the transition should have the same fitness. Unfortunately, pure neutrality transitions are likely to be rare among the adaptations that viruses undergo, and the data necessary to unambiguously identify them would be tedious if not impossible to obtain.

Our simulations provide evidence of selection for mutational robustness occurring through increasing the degree of neutrality of RNA sequences at population sizes far below the size of the neutral network that the sequences inhabit. Such increases in the degree of neutrality was recently found in a study that compared evolved RNA sequences to those deposited in an aptamer database (34). For example, the comparison showed that human tRNA sequences were significantly more neutral, and hence more robust to mutations, than comparable random sequences that had not undergone evolutionary selection. However, we must caution that while in our simulations selection for mutational robustness is the only force that can cause the sequences to become more mutationally robust, in real organisms other forces, for example selection for increased thermodynamic stability (1), could have similar effects.

An experimental system that is quite similar to our simulations, probably more so than typical RNA viruses, is that of *viroids*—unencapsidated RNA sequences of only around 300 bases—capable of infecting plant hosts. Viroid evolution appears to be limited by the need to maintain certain

secondary structural aspects (32), which is consistent with our fitness assumptions. Furthermore, in potato spindle tuber viroid (PSTVd), a wide range of single and double mutants are observed to appear after a single passage (37), suggesting that a quasispecies rapidly forms under natural conditions. Viroids may have agricultural applications as they are capable of inducing (desirable) dwarfism in certain plant species (27), and as such, a better understanding of their evolutionary processes may help to direct future research efforts.

Making the case for or against quasispecies dynamics in realistic, evolving populations of RNA viruses, or even just self-replicating RNA molecules, is not going to be easy. As the presence of an error threshold ((45, 47); see also discussion in (51)) or the persistence of a consensus sequence (this work) have been ruled out as a diagnostic, we have to look for markers that are both unambiguous and easy to obtain. Selection for robustness may eventually be observed in natural populations of adapting RNA viruses or viroids, but up to now, no such signals have been reported. Thus, while we can be confident that small population sizes do not preclude quasispecies dynamics in RNA virus populations, on the basis of current experimental evidence we cannot decide whether quasispecies selection takes place in RNA viruses.

2.6 Conclusions

Quasispecies effects are not confined to deterministic systems with infinite population size, but are readily observed in finite—even small—populations undergoing genetic drift. We find a continuous transition from very small populations, whose dynamics are dominated by drift, to larger populations, whose dynamics are dominated by quasispecies effects. The crucial parameter is the product of effective population size and genomic mutation rate, which needs to be significantly larger than one for quasispecies selection to operate. However, experimental evidence for these theoretical findings is currently not available, and will most likely be hard to obtain, because the differences in the dynamics of populations that are simply drifting and populations that are under quasispecies selection can be quite subtle. Thus, a dedicated experimental effort is needed to demonstrate quasispecies selection in natural systems.

Bibliography

[1] Bloom, J. D., Silberg, J. J., Wilke, C. O., Drummond, D. A., Adami, C., & Arnold, F. H. (2005). Thermodynamic prediction of protein neutrality. *Proc. Natl. Acad. Sci. USA*, *102*, 606–611.

[2] Bornberg-Bauer, E., & Chan, H. S. (1999). Modeling evolutionary landscapes: Mutational stability, topology, and superfunnels in sequence space. *Proc. Natl. Acad. Sci. USA*, *96*, 10689–10694.

[3] Burch, C. L., & Chao, L. (2000). Evolvability of an RNA virus is determined by its mutational neighbourhood. *Nature*, *406*, 625–628.

[4] Cowperthwaite, M., Bull, J. J., & Meyers, L. A. (2005). Distributions of beneficial fitness effects in RNA. *Genetics*, *170*.

[5] Demetrius, L., Schuster, P., & Sigmund, K. (1985). Polynucleotide evolution and branching processes. *Bull. Math. Biol.*, *47*, 239–262.

[6] Derrida, B., & Peliti, L. (1991). Evolution in a flat fitness landscape. *Bull. Math. Biol.*, *53*, 355–382.

[7] Domingo, E. (1992). Genetic variation and quasispecies. *Curr. Opin. Genet. Dev.*, *288*, 61–63.

[8] Domingo, E. (2002). Quasispecies theory in virology. *J. Virol.*, *76*, 463–465.

[9] Domingo, E., Biebricher, C. K., Eigen, M., & Holland, J. J. (2001). *Quasispecies and RNA Virus Evolution: Principles and Consequences*. Georgetown: Landes Bioscience.

[10] Domingo, E., & Holland, J. J. (1997). RNA virus mutations and fitness for survival. *Annu. Rev. Microbiol.*, *51*, 151–178.

[11] Domingo, E., Sabo, D., Taniguchi, T., & Weissmann, C. (1978). Nucleotide sequence heterogeneity of an RNA phage population. *Cell*, *13*, 735–744.

[12] Drake, J. W. (1993). Rates of spontaneous mutation among RNA viruses. *Proc. Natl. Acad. Sci. USA*, *90*, 4171–4175.

[13] Drake, J. W., & Holland, J. J. (1999). Mutation rates among RNA viruses. *Proc. Natl. Acad. Sci. USA*, *96*, 13910–13913.

[14] Eigen, M. (1971). Self-organization of matter and the evolution of macromolecules. *Naturwissenschaften*, *58*, 465–523.

[15] Eigen, M. (1996). On the nature of virus quasispecies. *Trends Microbiol.*, *4*, 216–218.

[16] Eigen, M., & Schuster, P. (1979). *The Hypercycle—A Principle of Natural Self-Organization*. Berlin: Springer-Verlag.

[17] Fontana, W., & Schuster, P. (1998). Continuity in evolution: on the nature of transitions. *Science*, *280*, 1451–1455.

[18] Fontana, W., & Schuster, P. (1998). Shaping space: the possible and the attainable in RNA genotype-phenotype mapping. *J. Theor. Biol.*, *194*, 491–515.

[19] Fontana, W., Stadler, P. F., Bornberg-Bauer, E. G., Griesmacher, T., Hofacker, I. L., Tacker, M., Tarazona, P., Weinberger, E. D., & Schuster, P. (1993). RNA folding and combinatory landscapes. *Phys. Rev. E*, *47*, 2083–2099.

[20] Forst, C. V., Reidys, C., & Weber, J. (1995). Evolutionary dynamics and optimization: Neutral networks as model-landscape for RNA secondary-structure folding-landscapes. In F. Morán, A. Moreno, J. J. Merelo, & P. Chacón (Eds.) *Advances in Artificial Life*, vol. 929 of *Lecture Notes in Artificial Intelligence*, (128–147). Springer.

[21] Grüner, W., Giegerich, R., Strothmann, D., Reidys, C., Weber, J., Hofacker, I. L., Stadler, P. F., & Schuster, P. (1996). Analysis of RNA sequence structure maps by exhaustive enumeration. 1. Neutral networks. *Monatshefte ur Chemie*, *127*, 355–374.

[22] Grüner, W., Giegerich, R., Strothmann, D., Reidys, C., Weber, J., Hofacker, I. L., Stadler, P. F., & Schuster, P. (1996). Analysis of RNA sequence structure maps by exhaustive enumeration. 1. Structures of neutral networks and shape space covering. *Monatshefte ur Chemie*, *127*, 375–389.

[23] Higgs, P. G., & Derrida, B. (1991). Stochastic models for species formation in evolving populations. *J. Phys. A*, *24*, L985–L991.

[24] Higgs, P. G., & Derrida, B. (1992). Genetic distance and species formation in evolving populations. *J. Mol. Evol.*, *35*, 454–465.

[25] Hofacker, I. L., Fontana, W., Stadler, P. F., Bonhoeffer, S., Tacker, M., & Schuster, P. (1994). Fast folding and comparison of RNA secondary structures. *Monatshefte f. Chemie*, *125*, 167–188.

[26] Holmes, E. C., & Moya, A. (2002). Is the quasispecies concept relevant to RNA viruses? *J. Virol.*, *76*, 460–462.

[27] Hutton, R. J., Broadbent, P., & Bevington, K. B. (2000). Viroid dwarfing for high density citrus plantings. *Hortic. Rev.*, *24*, 277–317.

[28] Huynen, M. A. (1996). Exploring phenotype space through neutral evolution. *J. Mol. Evol.*, *43*, 165–169.

[29] Huynen, M. A., & Hogeweg, P. (1994). Pattern generation in molecular evolution: Exploitation of the variation in RNA landscapes. *J. Mol. Evol.*, *39*, 71–79.

[30] Huynen, M. A., Stadler, P. F., & Fontana, W. (1996). Smoothness within ruggedness: The role of neutrality in adaptation. *Proc. Natl. Acad. Sci. USA*, *93*, 397–401.

[31] Jenkins, G. M., Worobey, M., Woelk, C. H., & Holmes, E. C. (2001). Evidence for the non-quasispecies evolution of RNA viruses. *Mol. Biol. Evol.*, *18*, 987–994.

[32] Keese, P., & Symons, R. (1985). Domains in viroids: Evidence of intermolecular RNA rearrangements and their contribution to viroid evolution. *Proc. Natl. Acad. Sci. USA*, *82*, 4582–4586.

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

[34] Meyers, L. A., Lee, J. F., Cowperthwaite, M., & Ellington, A. D. (2004). The robustness of naturally and artificially selected nucleic acid secondary structures. *J. Mol. Evol.*, *58*, 681–691.

[35] Novella, I. S. (2004). Negative effect of genetic bottlenecks on the adaptability of vesicular stomatitis virus. *J. Mol. Biol.*, *336*, 61–67.

[36] Nowak, M. A., & Schuster, P. (1989). Error thresholds of replication in finite populations — Mutation frequencies and the onset of muller's ratchet. *J. Theor. Biol.*, *137*, 375–395.

[37] Owens, R. A., & Thompson, S. M. (2005). Mutational analysis does not support the existence of a putative tertiary structural element in the left terminal domain of *potato spindle tuber viroid*. *J. Gen. Virol.*, *86*, 1835–1839.

[38] Reidys, C., Forst, C. V., & Schuster, P. (2001). Replication and mutation on neutral networks. *Bull. Math. Biol.*, *63*, 57–94.

[39] Reidys, C., Stadler, P., & Schuster, P. (1997). Generic properties of combinatory maps: Neutral networks on RNA secondary structures. *Bull. Math. Biol.*, *59*, 339–397.

[40] Schuster, P., Fontana, W., Stadler, P. F., & Hofacker, I. L. (1994). From sequences to shapes and back: A case study in RNA secondary structures. *Proc. R. Soc. Lond. (Biol)*, *255*, 279–284.

[41] Schuster, P., & Swetina, J. (1988). Stationary mutant distributions and evolutionary optimization. *Bull. Math. Biol.*, *50*, 635–660.

[42] Steinhauer, D. A., de la Torre, J. C., Meier, E., & Holland, J. J. (1989). Extreme heterogeneity in populations of vesicular stomatitis virus. *J. Virol.*, *63*, 2072–2080.

[43] van Lint, J. H., & Wilson, R. M. (2001). *A course in combinatorics*. Cambridge: Cambridge University Press, 2nd edn.

[44] van Nimwegen, E., Crutchfield, J. P., & Huynen, M. (1999). Neutral evolution of mutational robustness. *Proc. Natl. Acad. Sci. USA*, *96*, 9716–9720.

[45] Wagner, G. P., & Krall, P. (1993). What is the difference between models of error thresholds and Muller's ratchet? *J. Math. Biol.*, *32*, 33–44.

[46] Walter, A. E., Turner, D. H., Kim, J., Lyttle, M. H., Müller, P., Mathews, D. H., & Zuker, M. (1994). Coaxial stacking of helixes enhances binding of oligoribonucleotides and improves predictions of RNA folding. *Proc. Natl. Acad. Sci. USA*, *91*, 9218–9222.

[47] Wiehe, T. (1997). Model dependency of error thresholds: The role of fitness functions and contrasts between the finite and infinite sites models. *Genet. Res.*, *69*, 127–136.

[48] Wilke, C. O. (2001). Adaptive evolution on neutral networks. *Bull. Math. Biol.*, *63*, 715–730.

[49] Wilke, C. O. (2001). Selection for fitness versus selection for robustness in RNA secondary structure folding. *Evolution*, *55*, 2412–2420.

[50] Wilke, C. O. (2004). Molecular clock in neutral protein evolution. *BMC Genetics*, *5*, 25.

[51] Wilke, C. O. (2005). Quasispecies theory in the context of population genetics. *BMC Evol. Biol.*, *5*, 44.

[52] Wilke, C. O., & Adami, C. (2001). Interaction between directional epistasis and average mutational effects. *Proc. Roy. Soc. London Ser. B*, *268*, 1469.

[53] Wilke, C. O., & Adami, C. (2003). Evolution of mutational robustness. *Mutat. Res*, *522*, 3–11.

[54] Wilke, C. O., Lenski, R. E., & Adami, C. (2003). Compensatory mutations cause excess of antagonistic epistasis in RNA secondary structure folding. *BMC Evol. Biol.*, *3*, 3.

[55] Wilke, C. O., Wang, J. L., Ofria, C., Lenski, R. E., & Adami, C. (2001). Evolution of digital organisms at high mutation rate leads to survival of the flattest. *Nature*, *412*, 331–333.

[56] Wolf, J. B., Brodie, E. D., & Wade, M. J. (2000). *Epistasis and the Evolutionary Process*. Oxford: Oxford University Press.

Chapter 3

Quasispecies in Time-Dependent Environments

The work presented here is heavily based on:

C. O. Wilke, R. Forster and I. S. Novella, Quasispecies in time-dependent environments. *Current Topics in Microbiology and Immunology*, 299 (2006). Reprinted with copyright permissions.

3.1 Abstract

In recent years, quasispecies theory in time-dependent (that is, dynamically changing) environments has made dramatic progress. Several groups have addressed questions such as how the time scale of the changes affect viral adaptation and quasispecies formation, how environmental changes affect the optimal mutation rate, or how virus and host coevolve. Here, we review these recent developments, and give a nonmathematical introduction to the most important concepts and results of quasispecies theory in time-dependent environments. We also compare the theoretical results with results from evolution experiments that expose viruses to successive regimes of replication in two or more different hosts.

3.2 Introduction

From the point of view of a virus, the world is in a state of constant change. For sustained existence, the virus has to move from host organism to host organism, and each new host differs somewhat from the previous one. Some viruses even change their host species on a regular basis. For example, arboviruses alternate between infections in arthropods and infections in vertebrates. Within a host, the world is not necessarily constant either. Viruses infecting higher organisms find themselves under constant attack from an adaptive immune system, whose ability to fight any particular virus mutant grows dramatically after exposure to the mutant. Fortunately—for the virus, that is— the large class of RNA viruses are more than prepared for this world of change. Their high replication rates, combined with exceptionally high mutation rates, allow them to explore vast numbers of mutations, which leads to rapid adaptation in the face of environmental variation. For example, it has been estimated that HIV-1 produces all possible single-point mutants and a good fraction of the double mutants in every single patient every day (37).

The theoretical framework in which RNA viruses are typically modeled is quasispecies theory, which was put forward by Eigen and co-workers in the 1970s (14, 12, 13, 3). Quasispecies theory describes how a virus population reacts to the two forces selection and mutation, and how these two

forces counterbalance each other. Mathematically, quasispecies theory is equivalent to the theory of mutation-selection balance, which has been developed in parallel to quasispecies theory in the population-genetics literature (6).

Eigen and co-workers (14, 12, 13) defined the quasispecies as the stable mutant configuration that arises when mutation and selection are in perfect equilibrium. With this definition, the theory seemed to apply only to static enviroments and an infinite population size. Indeed, it has been argued that since viral populations are finite, quasispecies theory may not realistically reflect their behavior, because perfect equilibrium between mutation and selection can never be achieved in the real world (19, 18). However, a number of theoretical studies have shown that understanding the infinite-population dynamics is crucial for developing a more detailed finite-population model (42, 44), and that all the effects predicted by quasispecies theory can in principle be observed also in finite populations (33, 41, 50, 45).

In recent years, substantial progress has been made in extending quasispecies theory to dynamic environments. This work can be subdivided into two groups, studies on virus evolution in time-dependent fitness landscapes (27, 28, 29, 49, 48) and studies on the interaction between virus and the host's immune system, or virus-host coevolution in general (21, 22, 5). In this chapter, we put most emphasis on the former type of studies, but will also give a brief overview over the latter. We discuss the relationship of theory to experimental results in the individual sections where appropriate.

3.3 Virus Evolution in Time-Dependent Fitness Landscapes

In this section, we consider situations in which the virus's environment changes due to external factors, but is not in return influenced by any of the virus's adaptive moves. Throughout this chapter, we use for this case the terminology of adaptation to a time-dependent fitness landscape. As an example, consider a virus that is grown alternatingly on different hosts (see also section 3.4).

From a theoretical point of view, it is particularly interesting to study changes in the fitness landscape that are periodic (that is, situations in which the environment changes over and over again in exactly the same pattern), because the theory that has been developed for static fitness

landscapes can—after a suitable mathematical transformation—to a large extent be directly applied to periodic fitness landscapes (49). A number of exact results and simple principles have been derived for periodic fitness landscapes. We will review these results in the following paragraphs, and will consider nonperiodic changes briefly at the end of this section.

Before we can talk in depth about periodically changing (or oscillating) fitness landscapes, we have to introduce some terminology. We refer to the time it takes for the fitness landscape to go through one cycle of changes as the oscillation period. For example, if the fitness landscape alternates between environment A for 100 generations and environment B for 100 generations, then the oscillation period is 200 generations. We denote the oscillation period by T. The inverse of the oscillation period, 1/T, is called the *oscillation frequency*. The shorter the oscillation period the higher the oscillation frequency. We refer to a special point in time within one oscillation period as the phase. We typically express the phase in fractions of one complete oscillation cycle. For example, in the above example, the switch from A to B occurs at phase 1/2, and the switch from B to A occurs at phase 1, which by definition is equivalent to phase 0.

One important result for oscillating fitness landscapes is that as long as we look at the system under study only at a single phase (say, for example, at the beginning of the oscillation, phase = 0), the virus population behaves as if the fitness landscape were static (49). After an initial transient period the virus population develops a stable distribution of mutants, a quasispecies. The only difference to a static fitness landscape is that this equilibrium depends on the phase. At phase 1/2, the equilibrium distribution of mutants may be totally different from the distribution at phase 0, but at any given phase, the equilibrium distribution will always be the same. In theory, the population reaches this equilibrium only after infinitely many oscillation cycles have passed, but in practice, more than a couple of hundred cycles are rarely necessary to obtain equilibration, and frequently equilibration occurs after only a handful of cycles.

We can also calculate accurately what happens when the oscillations are very fast or very slow. The general result is the following: Under very fast changes, the population behaves as if the fitness landscape did not change at all, but instead were the average of the changing landscape (49, 48).

That is, if for example two traits are alternatingly selected for, then above a certain oscillation frequency, the population will simply acquire both traits at the same time. For very slowly changing landscapes, the population essentially does not realize that the environment is changing, and climbs the local peak that best represents the current environment. If the environment changes slowly but continuously, then the population will track this local optimum, and always stay optimally adapted (29). If however the environment suddenly requires a completely different adaptation, then the population may be thrown off track and suddenly find itself hopelessly ill adapted and unable to cope with the new environment (38).

The behavior at intermediate oscillation frequencies falls between the two extremes. As the oscillation frequency goes up, the population ceases to track the changes accurately, and behaves more as if it were only experiencing the average fitness landscape. An interesting phenomenon that occurs at intermediate oscillation frequencies is the phase shift. The term phase shift means that there is a substantial lag between the point in time at which a change in the fitness landscape takes place and the point in time at which the population shows response to this change. If we measure this lag in units of the oscillation period, then the lag goes to zero as the oscillation period increases. However, when the oscillation period decreases, then the phase shift can assume substantial values, as the population is constantly trying to keep up with the changes in the landscape (29). To summarize, we can compare the evolutionary dynamics in a periodic fitness landscape with the effect of a low-pass filter on an audio signal (49, 29): A low-pass filter replaces high-frequency signals with their average intensity, and lets low-frequency signals pass unaltered. Moreover, for intermediate to high frequency signals, a low-pass filter introduces a time lag (phase shift) that depends on the frequency of the signal.

The general predictions for a fitness landscape whose changes are aperiodic or irregular are very similar to those for a periodically changing fitness landscape. Again, evolution essentially acts as a low-pass filter on the changing fitness landscape. The population will not be able to follow changes in the fitness landscape that happen on a short timescale, and will see only the average effect of these changes. However, the population will typically be able to track accurately a slowly changing

landscape and stay optimally adapted. One significant difference to periodic fitness landscapes is the case of very fast, irregular changes. In this case, it can happen that the population loses track of the advantageous regions in the fitness landscape altogether, and becomes completely ill adapted. This loss of adaptation can be considered as the dynamic equivalent of the classic (static) error threshold. The static error threshold sets an upper bound to the maximum mutation rate that a population can sustain without losing all genetic information (14). The dynamic error threshold differs from the classic error threshold in that it sets a lower bound on the minimum mutation rate with which a population can still follow the changes in the fitness landscape (27).

3.4 Adaptation to Two Alternating Hosts

As an example for a periodic fitness landscape, we consider here the situation where a virus population has to adapt to two alternating hosts. Such a situation is common for example in arboviruses, which alternate between rounds of replication in arthropod vectors and rounds of replication in vertebrate hosts. One question that is frequently asked in this context is whether the virus can adapt to both hosts at the same time, or whether adaptation to one host is typically accompanied by loss of adaptation to the other host. We can study this question in the framework of a very simple model. Assume that a virus strain has one gene that is advantageous in host one and neutral in host two, and a second gene that has the reversed properties. We refer to this strain as the divided strain. Further, assume that there is also a second virus strain which has a single gene that conveys the advantageous functions of each of the two genes in the corresponding hosts. We refer to this strain as the fused strain. In addition, both the divided and the fused strain can suffer from point mutations, and do so at the same rate per site. In the case of the divided strain, a single point mutation will affect either of the two genes (and thus the two functions) separately, whereas in the fused strain, point mutations affect both functions simultaneously. Note that while we talk about separate genes throughout this chapter, this model can also apply to separate functions carried out by a single gene. In this case, the divided strain corresponds to a gene that can adapt to either function, but not to both at the same time, wherease the fused strain corresponds to a gene that

can adapt to both functions at the same time. These differences in the strains can be caused by pleiotropic effects.

Throughout this section, we consider populations of finite size N, whereas the general results presented in the previous section are derived from infinite-population theory. By and large, these general results also apply to finite populations. The main difference is that (i) in finite populations, genotypes can truly go extinct, whereas all genotypes are always present in infinite populations, and (ii) that the quasispecies dynamic becomes a stochastic process, subject to fluctuations and chance events.

Figure 3.1 demonstrates how the divided strain will evolve if it is grown alternatingly for 30 generations each on one of the two hosts. When the divided strain grows on host one, the second gene is not under selection, and therefore slowly deteriorates due to mutation accumulation. When the virus is transferred to the second host, then gene two comes under selection and rebounds quickly, while gene one starts to deteriorate. Figure 3.2 illustrates this dynamic by visualizing the mutant cloud that forms at different points in time. When gene one is under selection, then only a small number of sequences acquire mutations in this gene, while at the same time the majority of the population acquires mutations in gene two. The image reverses when gene two starts to come under selection. Then, only a small number of sequences acquire mutations in gene two, but almost all sequences acquire mutations in gene one. Note that the fused strain, grown under the same conditions, fully retains both functions at all times, and accumulates only a small number of mutations in any case (fig. 3.2).

If we compete the divided with the fused strain in a changing environment, which one will prevail? There is no simple answer to this question. The answer depends on a number of factors, the two most important of which are the time scale on which the changes in the environment occur and whether mutation pressure creates any additional trade-offs between the divided strain and the fused type. For example, if the environment changes very quickly (such as when the host changes every generation), then the divided strain will maintain both functions at all times, just as the fused type does. Therefore, in this case the only difference between the two strains is the extent

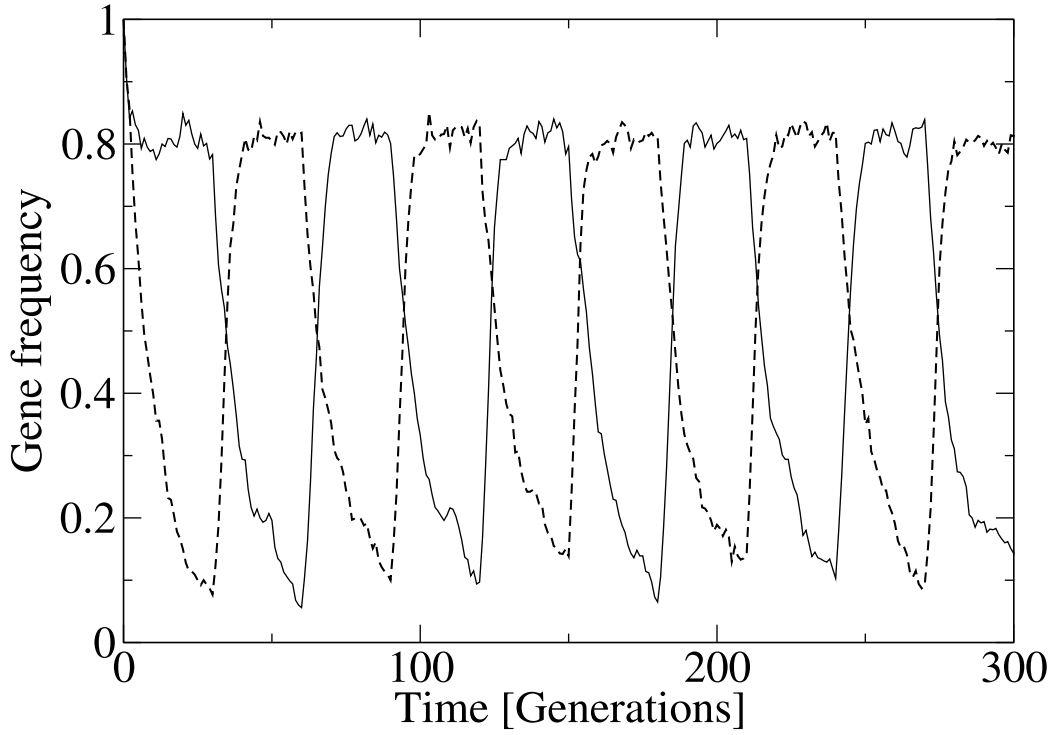


Figure 3.1: Frequencies of error-free genes as a function of time. Parameters are: Oscillation period $T = 60$, per-site mutation rate $\mu = 0.02$, length of a single gene of the divided strain $l_{\text{div}} = 5$, selective advantage of functional gene $s = 1$, population size $N = 1000$.

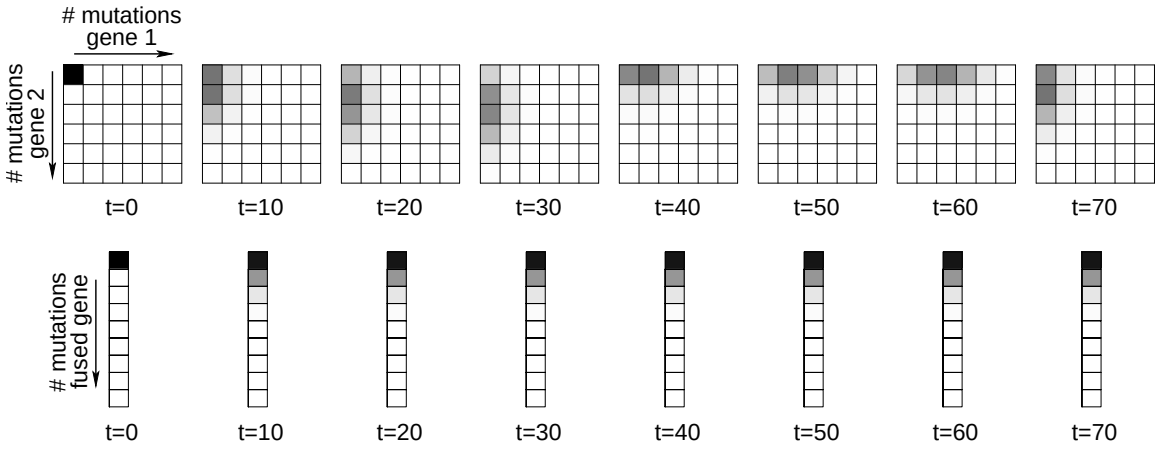


Figure 3.2: Population structure of either the divided or fused strain at various time points. Gray levels indicate the fraction of sequences at the given mutational distance from the respective error-free sequence. Parameters are: Oscillation period $T = 60$, per-site mutation rate $\mu = 0.02$, length of a single gene of the divided strain $l_{\text{div}} = 5$, length of the fused gene $l_{\text{fused}} = 8$, selective advantage of functional gene $s = 1$, infinite population size.

to which they suffer from mutation pressure. If we assume that their per-site mutation rate μ is the same, then the mutation pressure translates into the number of sites in the genome that are under selection. The strain with the larger number of sites under selection will suffer from a higher mutational pressure, and lose against the other strain (47). Thus, if the fused gene can perform both functions with fewer sites under selection than the two genes of the divided strain, then the fused strain has an advantage. A similar argument applies when the environment changes very slowly. However, in this case the selection pressure is different, because the virus has sufficient time to adapt fully to one host before it is exposed again to the other host. During this long period of exposure to a single host, the decisive factor is now whether the single gene of the divided strain that is needed for this host suffers from a higher mutational pressure than the fused gene.

It is probably reasonable to assume that the two genes of the divided strain each have fewer selective sites than the single fused gene, but that the fused gene has fewer selective sites than the two former genes together. Under this assumption, the fused strain will be advantageous when the environment changes quickly, while the divided strain will be advantageous when the environment changes slowly. Figure 3.3 shows how the fused strain can invade a population of the divided strain if the environmental changes occur on a sufficiently fast time scale, while fig. 3.4 shows how the divided strain can invade a population of the fused strain if the environment changes slowly.

In experiments with digital organisms (self-replicating computer programs that mutate and evolve, (46)) that were adapting to a periodically changing fitness landscape, trajectories very similar to the one shown in fig. 3.3 were observed for moderately to rapidly changing fitness landscapes, but not for very slowly changing fitness landscapes (26). These observations imply that genes or part of genes can indeed evolve to be fused or divided, even if the system under study does not *a priori* provide fused or divided genes. In the digital organisms, separation or fusion of genes was not imposed from the outside, unlike in the case of the simple toy model discussed in this section, and evolved presumably through changes in the amount of epistatic interactions among the different parts of the digital organisms' genomes.

Several groups have experimentally tested the effect of replication in two different hosts on the

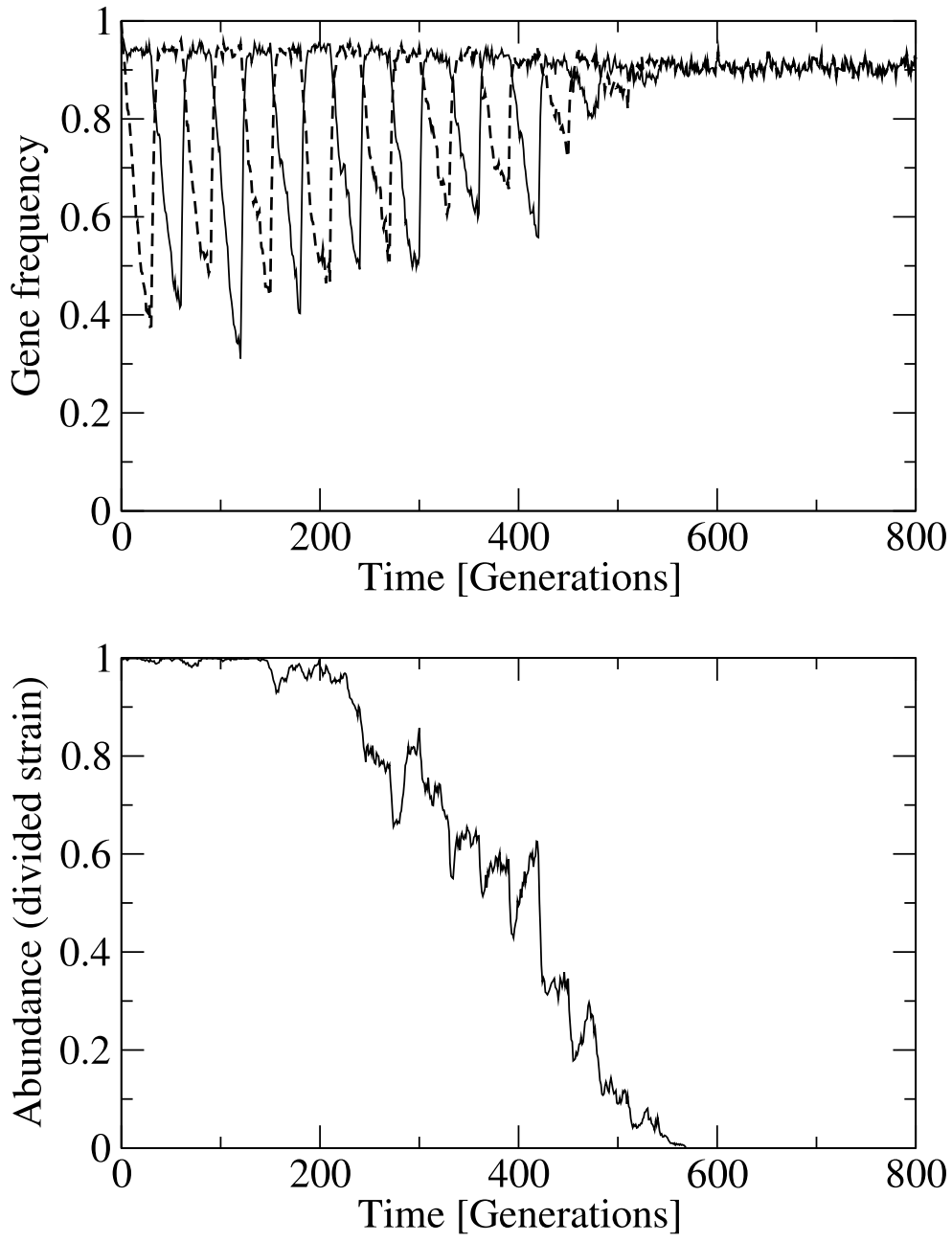


Figure 3.3: Fused strain invades a population of the divided strain. A single fused genotype is introduced into the population at $t = 0$, and reintroduced at later times whenever all fused genotypes have been lost from the population. The 25th introduced genotype goes to fixation. Parameters are: Oscillation period $T = 60$, per-site mutation rate $\mu = 0.006$, length of a single gene of the divided strain $l_{\mathrm{div}} = 5$, length of the fused gene $l_{\mathrm{fused}} = 8$, selective advantage of functional gene $s = 1$, population size $N = 1000$.

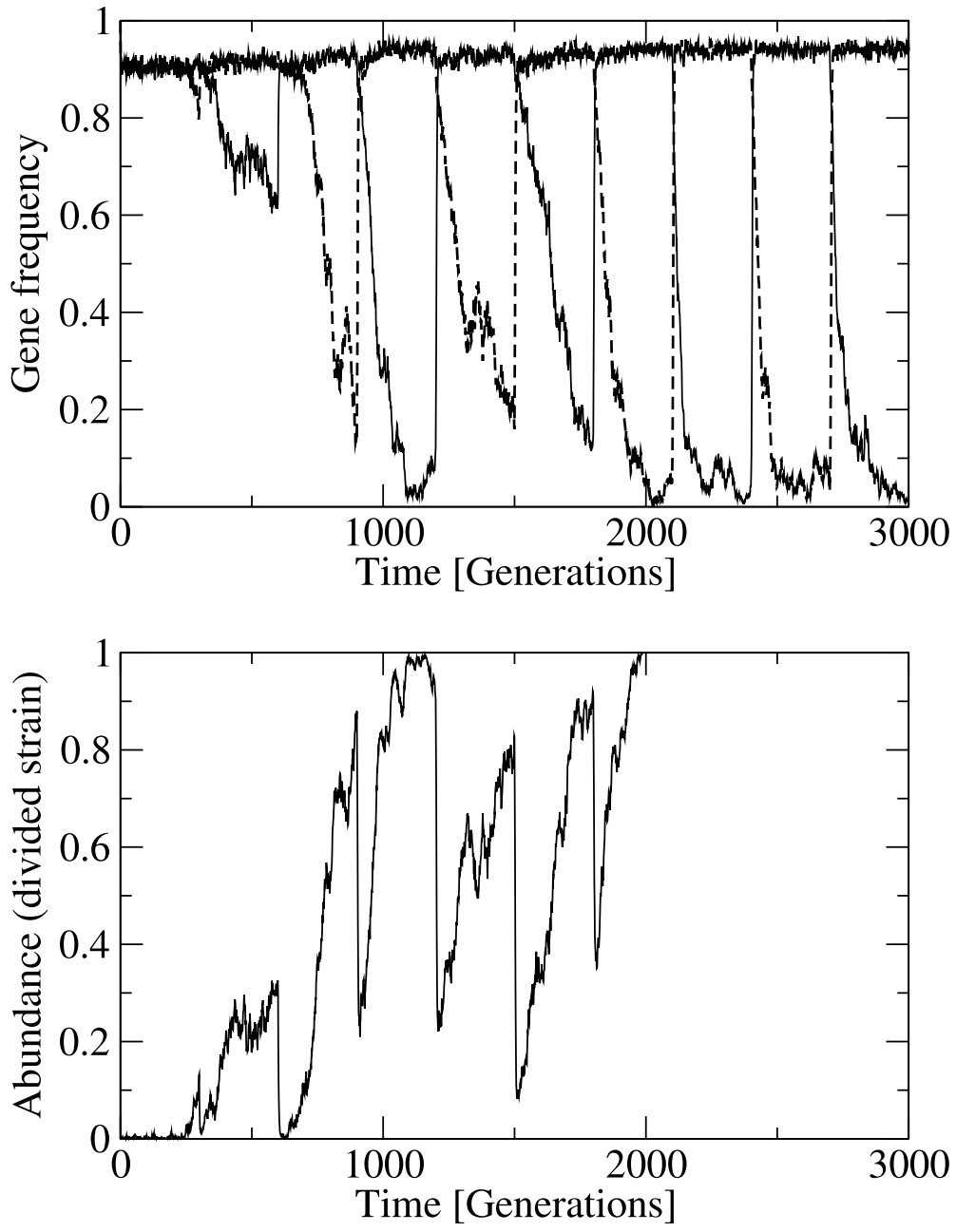


Figure 3.4: Divided strain invades a population of the fused strain. A single divided genotype is introduced into the population at $t = 0$, and reintroduced at later times whenever all divided genotypes have been lost from the population. The 31st introduced genotype goes to fixation. Parameters are: Oscillation period $T = 600$, per-site mutation rate $\mu = 0.006$, length of a single gene of the divided strain $l_{\mathrm{div}} = 5$, length of the fused gene $l_{\mathrm{fused}} = 8$, selective advantage of functional gene $s = 1$, population size $N = 1000$.

evolution of viral quasispecies. Alternating acute replication of vesicular stomatitis virus (VSV), eastern equine encephalitis virus (EEEV), and Dengue virus in cells of insect and mammalian origin led to fitness increase in both cells types (31, 43, 8). The difference in behavior between VSV and EEEV is that for VSV, insect and mammalian cells represent similar fitness landscapes for replication, whereas for EEEV a trade-off can be observed: Adaptation to one cell type results in loss of viral fitness in the other. Similar results are observed for adaptation to cells of insect and avian origin. Viruses that grow alternatingly on these two cell types produce adaptations to both environments in spite of trade-offs (9). However, while acute insect replication does not involve trade-off for acute mammalian replication, persistent replication of VSV in insect cells does result in trade-off for replication in mammalian cells (30), and evolution during alternation between acute mammalian replication and persistent insect replication is dominated by the persistent phase (51). Additional work has been done using different types of mammalian cells (BHK-21, MDCK and HeLa) as hosts for VSV replication. In these experiments, VSV consistently adapted to both of the alternating cell types independently of the existence of trade-offs and independently of whether the host switch occurred periodically or randomly (40). All these experimental results are in qualitative agreement with the theory, in that they show that there is no simple rule of how viruses behave that are grown alternatingly in two different hosts. They may either adapt to both hosts at the same time, or adapt to one host at the expense of adaptation to the other. Which of the two cases occurs depends on the time spent in each of the two hosts, and also on the details of the fitness landscape in the two hosts.

3.5 Adaptation of Mutation Rate

In the context of time-dependent environments, it is natural to ask what is the optimal mutation rate for the viral population. In a static environment, once a virus has reached the global optimum in the fitness landscape, all mutations are deleterious. In this case, it pays for the virus to reduce the mutation rate as much as possible, as a lower mutation rate translated into a lower mutational load—and thus higher average fitness—for the offspring virions. In a time-dependent environment,

on the other hand, a zero or near-zero mutation rate is detrimental. A virus needs a sufficiently high mutation rate in order to be able to follow changes in the fitness landscape. It is clear that the mutation rate at which the virus is optimally capable of tracking the fitness landscape will depend on the speed and nature of the changes in the fitness landscape.

Nilsson and Snoad (28) studied a simple model of a viral quasispecies that has to track a moving fitness peak, and calculated the optimal mutation rate in this model. The model assumptions are the following: The fitness peak remains static for a time interval of length τ_1. Then, it moves to a position nearby in the sequence space. The distance from the old position to the new position is k_1 point mutations. The peak remains static at the new position for a time interval of length τ_2, until it moves to another nearby position, this time with distance k_2, and so on. The lengths of the intervals τ_i and the distances k_i need not be identical from shift to shift. Nilsson and Snoad found that the optimal mutation rate could be expressed simply as $\mu_{\mathrm{opt}} = \langle k \rangle / (\langle \sigma \rangle \langle \tau \rangle)$, where $\langle k \rangle$ is the average mutational distance that the the peak moves, $\langle \sigma \rangle$ is the mean replication rate of the virus on the peak, and $\langle \tau \rangle$ is the average time between peak shifts.

We can rewrite the result of Nilsson and Snoad such that its interpretation becomes more intuitive. Note that $1/\langle \sigma \rangle$ is the average time the virus needs to replicate, and that therefore $\langle \sigma \rangle \langle \tau \rangle$ is the average time between peak shifts measured in virus generations. Thus Nilsson and Snoad's result states that the virus can optimally adapt to the changing environment if its genomic mutation rate per generation is the mean distance between peaks divided by the number of generations between successive peak shifts. This result makes intuitive sense: With this mutation rate, the majority of mutants in the virus population will by the next peak shift have accumulated exactly the right number of mutations to be located at the position of the new peak. It is interesting to note that for $\langle k \rangle = 1$, we recover a result that Kimura had already derived in a much simpler model in 1967 (23).

Note that Nilsson and Snoad's result is very different from the optimal mutation rate of an ill adapted virus in a static environment, that is, of a virus that can still discover many advantageous mutations even when the environment does not change. In this case, the optimal mutation rate is given by the mean effect of deleterious mutations (35). A mutation rate of this magnitude optimally

balances the beneficial effect of new advantageous mutations with the deleterious effect of too high a mutational load. It is not clear whether Orr's theory or Nilsson and Snoad's theory is more relevant to RNA viruses. The answer depends on the time scale on which changes happen, and on how easy it is for a virus to reach optimal adaptation after a change in the fitness landscape has taken place. If changes happen only occasionally, and a number of beneficial mutations need to be discovered to reach optimal adaptation to the new environment, then Orr's theory is probably more relevant. If on the other hand changes happen regularly, and only one or two mutations are needed to reach optimal adaptation to the new environment, then Nilsson and Snoad's theory should apply.

3.6 Coevolution

In the preceeding sections, we have discussed virus evolution in fitness landscapes that change because of external forces. By "external forces" we mean that the virus has to adapt to environmental changes, but the environment is not influenced by the adaptive moves of the virus. In many cases, however, the virus imposes a time-dependent fitness landscape on the environment just as much as the environment imposes a time-dependent fitness landscape on the virus. Typical examples are the interaction between virus and immune system, and coevolution of virus and host in general. This section is not meant as an in-depth review of these areas, but rather as an entry point that should help the reader locate the relevant literature in this field.

Kamp and co-workers (21, 22) have extended earlier work on a virus population that is tracking a moving optimum (27) to the case where the moving optimum is defined by an adaptive immune system, which itself is tracking the virus population (see also (20) for a review). Interestingly, in this case not only does the virus population form a quasispecies, but the immune system can be interpreted as a quasispecies as well: Under B cell proliferation and somatic hypermutation, a mutant cloud of B cells developes that behaves similarly to a viral quasispecies. Kamp and Bornholdt found that both the viral quasispecies and the B cell quasispecies suffer from the classical error threshold if their mutation rates are too high. In addition, the viral quasispecies suffers also from a dynamic threshold if its mutation rate is too low in comparison to that of the B cells. In this case, the virus

cannot escape rapidly enough from the immune system, and is driven to exinction. Kamp and co-workers also calculated the optimum mutation rates for both the immune system and the virus, and found that both results compare favorably with experimental observations (21, 22). The optimal mutation rate for the virus has a particularly simple and intuitive interpretation: The mutation rate should be such that each viral epitope acquires on average exactly one mutation within the time frame in which the immune system reacts to that epitope. (Note that this result is essentially identical to the one of Kimura (23) discussed in the preceeding section.)

Brumer and Shakhnovich (5) have extended this work to study the effect of semiconservative replication (where both template and copied sequence acquire mutations, as is the case in replication of double-stranded DNA) versus conservative replication (where mutations arise only in the copied sequence, as is the case with most RNA-based viruses). They have found that the optimum viral mutation rate is largely unaffected by the mode of replication, but that the model dynamics away from this optimum differ substantially with the mode of replication. Brumer and Shakhnovich suggest that their result might explain why DNA-based viruses tend to have much smaller mutation rates than RNA based viruses.

In general, a substantial amount of modeling work on the interaction between viruses and immune systems has been done by theoretical immunologists. While many of the basic immunological models do not explicitly consider evolution, they provide a natural starting point for more realistic models that do include evolutionary effects. Good introductions to theoretical immunology and its relation to virus dynamics and virus evolution are the book by Nowak and May (32) and a recent review article by Perelson (36).

Coevolution of viruses and host cells has been studied experimentally in several viral species, particularly during persistent infection. Infection of BHK-1 cells with foot-and-mouth-disease virus can select cells of increased resistance, which in turn selects viral populations of increased virulence for nonselected host cells (25, 16). Similar results have been reported for experiments with persistent infection of poliovirus (4), reovirus (1, 11), and mouse hepatitis virus (7). As a consequence, persistently infected cells are less susceptible to viral infection, and this susceptibility tends to correlate

with the ability (or lack thereof) of the virus to enter the cells (4, 11, 7).

Another model of coevolution that has been studied extensively both theoretically and experimentally is that of defective interfering particles (DIPs) and their full-length wild-type counterparts. DIPs are mutant viruses with large deletions in their genomes. Because of these deletions, DIPs cannot produce all the proteins needed for replication, but they can use these proteins from coinfecting full-length wild-type helper viruses.

Several authors have studied the dynamics of DIPs and their associated helper viruses theoretically (2, 39, 24, 15). DIPs can accumulate when the multiplicity of infection (m.o.i.) is high, for two reasons: First, at high m.o.i., DIPs frequently coinfect cells together with the wild type, and thus have access to the wild type proteins that they need for replication. Second, DIPs have a smaller genome size than the wild type, and thus can replicate faster than the wild type and use up a larger share of wild-type protein than the wild type does itself. The simplest models of this system predict that DIPs and the wild type will evolve towards a stable equilibrium, and that a DIP-resistant mutant can occur which will replace the DIP-wild type system (39). More elaborate models that take into account the changing size of the virus population, for example under recurrent passaging, show that fluctuating levels of DIPs and wild type are more realistic, and that the DIP-wild type system can spontaneously go to extinction (2, 24, 15). For example, extinction happens when a severe bottleneck removes all wild-type particles from the virus population.

While DIPs have been identified in many viral species (reviewed in (17)), persistent VSV infection of mammalian cells is the system of DIP-driven evolution that is characterized in most detail. DIPs accumulate during infection of mammalian cells at high m.o.i., but not at low m.o.i. Eventually, DIP-resistant mutant genomes of full length appear, and rise to fixation as they replace both the wild type and DIPs. These mutants are typically characterized by extensive rearrangement of the genomic termini (34). If viral replication continues at high m.o.i., then new DIPs arise that can use the full-length mutant genomes as helper viruses. Thus starts a second phase of interference. Such cycles of the emergence of interfering DIPs followed by the emergence of resistant full-length genomes can proceed for long periods of time (10).

3.7 Discussion

While the first 25 years of quasispecies theory were dominated by studies of static fitness landscapes, in recent years several groups have started to analyze quasispecies dynamic in time-dependent environments. It is comforting to see that by and large, the results derived for static fitness landscapes apply—with only minor modifications—to dynamic fitness landscapes as well. There is no need to replace quasispecies theory with some other theory just because viruses do not live in perfectly stable environments. On the contrary, the theory for periodically changing fitness landscapes, for example, is actually mathematically equivalent to that of a static fitness landscape, and can be analyzed with exactly the same tools.

However, dynamic fitness landscapes do, of course, introduce new effects that cannot be observed in static fitness landscapes. The most dramatic such effect is probably the dynamic error threshold, discovered by Nilsson and Snoad (27), which puts a lower bound on the minimum mutation rate at which a virus population is still capable of following the changes in the landscape. This dynamic error threshold, in combination with the classical error threshold known from static landscapes, restricts the useful range of mutation rates to a fairly narrow region, and it has been suggested that the mutation rates of several RNA viruses fall right into this region (21, 22).

A particularly useful way to describe the overall behavior of an evolving population on a dynamic fitness landscape is to say that the population acts like a low-pass filter. The slower the changes in the landscape, the more accurately the population will track them, and the smaller the time lag from the change in the landscape until the population reacts to it. The faster the changes, on the other hand, the less accurately the population will track them, and the larger the time lag until the population reacts. In the limit of very fast changes, the population will stop to track the changes completely, and simply adapt to the time-averaged fitness landscape.

This review has shown that we have a fairly good understanding of quasispecies theory in time-dependent environments, at least on a broad, qualitative level. However, as always, the devil is in the detail, and there remain many open questions. Most importantly, the theoretical models tend to be based on simplified fitness landscapes, in which a single gene sequence provides full function,

and any mutation to that sequence destroys or at least severely inhibits the function. However, we know that mutations can come with a broad spectrum of effects, from fully neutral to somewhat deleterious to lethal. In a changing environment, this spectrum may change over time as well, for example as a function of the virus host. One of the main goals of future work in this area must be to determine the detailed structure of realistic fitness landscapes and how exactly they change over time. To realize this goal, close collaboration between theoretical and experimental groups will be necessary.

Bibliography

[1] Ahmed, R., Canning, W. M., Kauffman, R. S., Sharpe, A. H., Hallum, J. V., & Fields, B. N. (1981). Role of the host cell in persistent viral infection — Coevolution of L cells and reovirus during persistent infection. *Cell*, *25*, 325–332.

[2] Bangham, C. R. M., & Kirkwood, T. B. L. (1990). Defective interfering particles: effects in modulating virus growth and persistence. *Virology*, *179*, 821–826.

[3] Biebricher, C. K., & Eigen, M. (2006). What is a quasispecies? *Curr. Top. Microbiol. Immunol.*, *299*, 1–32.

[4] Borzakian, S., Couderc, T., Barbier, Y., Attal, G., Pelletier, I., & Colberegarapin, F. (1992). Persistent poliovirus infection — Establishment and maintenance involve distinct mechanisms. *Virology*, *186*, 398–408.

[5] Brumer, Y., & Shakhnovich, E. I. (2004). Host-parasite coevolution and optimal mutation rates for semiconservative quasispecies. *Phys. Rev. E*, *69*, 061909.

[6] Bürger, R. (2000). *The mathematical theory of selection, recombination, and mutation*. Chichester: Wiley.

[7] Chen, W., & Baric, R. S. (1996). Molecular anatomy of mouse hepatitis virus persistence: coevolution of increased host cell resistance and virus virulence. *J. Virol.*, *70*, 3947–3960.

[8] Chen, W.-J., Wu, H.-R., & Chiou, S.-S. (2003). E/NS1 modifications of dengue 2 virus after serial passages in mammalian and/or mosquito cells. *Intervirology*, *46*, 289–295.

[9] Cooper, L. A., & Scott, T. W. (2001). Differential evolution of eastern equine encephalitis virus populations in response to host cell type. *Genetics*, *157*, 1403–1412.

[10] DePolo, N. J., Giachetti, C., & Holland, J. J. (1987). Continuing coevolution of virus and defective interfering particles and of viral genome sequences during undiluted passages — Virus mutants exhibiting nearly complete resistance to formerly dominant defective interfering particles. *J. Virol.*, *61*, 454–464.

[11] Dermody, T. S., Nibert, M. L., Wetzel, J. D., Tong, X., & Fields, B. N. (1993). Cells and viruses with mutations affecting viral entry are selected during persistent infections of L cells with mammalian reoviruses. *J. Virol.*, *67*, 2055–2063.

[12] Eigen, M., McCaskill, J., & Schuster, P. (1988). Molecular quasi-species. *J. Phys. Chem.*, *92*, 6881–6891.

[13] Eigen, M., McCaskill, J., & Schuster, P. (1989). The molecular quasi-species. *Adv. Chem. Phys.*, *75*, 149–263.

[14] Eigen, M., & Schuster, P. (1979). *The Hypercycle—A Principle of Natural Self-Organization*. Berlin: Springer-Verlag.

[15] Frank, S. A. (2000). Within-host spatial dynamics of viruses and defective interfering particles. *J. Theor. Biol.*, *206*, 279–290.

[16] Hernández, M. A. M. M., Carrillo, E. C., Sevilla, N., & Domingo, E. (1994). Rapid cell variation can determine the establishment of a persistent viral infection. *Proc. Natl. Acad. Sci. USA*, *91*, 3705–3709.

[17] Holland, J. J. (1991). Defective viral genomes. *Fundamental Virology*, (151–165).

[18] Holmes, E. C., & Moya, A. (2002). Is the quasispecies concept relevant to RNA viruses? *J. Virol.*, *76*, 460–462.

[19] Jenkins, G. M., Worobey, M., Woelk, C. H., & Holmes, E. C. (2001). Evidence for the non-quasispecies evolution of RNA viruses. *Mol. Biol. Evol.*, *18*, 987–994.

[20] Kamp, C. (2003). A quasispecies approach to viral evolution in the context of an adaptive immune system. *Microbes Infect*, *5*, 1397–1405.

[21] Kamp, C., & Bornholdt, S. (2002). Coevolution of quasispecies: B-cell mutation rates maximize viral error catastrophes. *Phys. Rev. Lett.*, *88*, 068104.

[22] Kamp, C., Wilke, C. O., Adami, C., & Bornholdt, S. (2002). Viral evolution under the pressure of an adapative immune response — Optimal mutation rates for viral escape. *Complexity*, *8*, 28–33.

[23] Kimura, M. (1967). On the evolutionary adjustment of spontaneous mutation rates. *Genet. Res.*, *9*, 23–34.

[24] Kirkwood, T. B. L., & Bangham, C. R. M. (1994). Cycles, chaos, and evolution in virus cultures: A model of defective interfering particles. *Proc. Natl. Acad. Sci. USA*, *91*, 8685–8689.

[25] la Torre, J. C. D., Martínez-Salas, E., Diez, J., Villaverde, A., Gebauer, F., Rocha, E., Dávila, M., & Domingo, E. (1988). Coevolution of cells and viruses in a persistent infection of foot-and-mouth disease virus in cell culture. *J. Virol.*, *62*, 2050–2058.

[26] Li, Y., & Wilke, C. O. (2004). Digital evolution in time-dependent fitness landscapes. *Artificial Life*, *10*, 123–134.

[27] Nilsson, M., & Snoad, N. (2000). Error thresholds on dynamic fitness landscapes. *Phys. Rev. Lett.*, *84*, 191–194.

[28] Nilsson, M., & Snoad, N. (2002). Optimal mutation rates in dynamic environments. *Bull. Math. Biol.*, *64*, 1033–1043.

[29] Nilsson, M., & Snoad, N. (2002). Quasispecies evolution on a fitness landscape with a fluctuating peak. *Phys. Rev. E*, *65*, 031901.

[30] Novella, I. S., Clarke, D. K., Quer, J., Duarte, E. A., Lee, C. H., Weaver, S. C., Elena, S. F., Moya, A., Domingo, E., & Holland, J. J. (1995). Extreme fitness differences in mammalian and

insect hosts after continuous replication of vesicular stomatitis virus in sandfly cells. *J. Virol.*, *69*, 6805–6809.

[31] Novella, I. S., Hershey, C. L., Escarmis, C., Domingo, E., & Holland, J. (1999). Lack of evolutionary stasis during alternating replication of an arbovirus in insect and mammalian cells. *J. Mol. Biol.*, *287*, 459–465.

[32] Nowak, M., & May, R. M. (2000). *Virus dynamics*. Oxford: Oxford Univerisity Press.

[33] Nowak, M. A., & Schuster, P. (1989). Error thresholds of replication in finite populations — Mutation frequencies and the onset of muller's ratchet. *J. Theor. Biol.*, *137*, 375–395.

[34] O'Hara, P. J., Nichol, S. T., Horodyski, F. M., & Holland, J. J. (1984). Vesicular stomatitis virus defective interfering particles can contain extensive genomic sequence rearrangements and base substitutions. *Cell*, *36*, 915–924.

[35] Orr, H. A. (2000). The rate of adaptation in asexuals. *Genetics*, *155*, 961–968.

[36] Perelson, A. S. (2002). Modelling viral and immune system dynamics. *Nat. Rev. Immunol.*, *2*, 28–36.

[37] Perelson, A. S., Essunger, P., & Ho, D. D. (1997). Dynamics of HIV-1 $CD4^+$ lymphocytse *in vivo*. *AIDS*, *11*, S17–S24.

[38] Ronnewinkel, C., Wilke, C. O., & Martinetz, T. (2001). Genetic algorithms in time-dependent environments. *Theoretical aspects of evolutionary computing*, (261–285).

[39] Szathmary, E. (1992). Natural selection and dynamical coexistence of defective and complementing virus segments. *J. Theor. Bio.*, *157*, 383–406.

[40] Turner, P. E., & Elena, S. F. (2000). Cost of host radiation in an RNA virus. *Genetics*, *156*, 1465–1470.

[41] van Nimwegen, E., Crutchfield, J. P., & Huynen, M. (1999). Neutral evolution of mutational robustness. *Proc. Natl. Acad. Sci. USA*, *96*, 9716–9720.

[42] van Nimwegen, E., Crutchfield, J. P., & Mitchell, M. (1999). Statistical dynamics of the royal road genetic algorithm. *Theo. Comp. Sci.*, *229*, 41–102.

[43] Weaver, S. C., Brault, A. C., Kang, W., & Holland, J. J. (1999). Genetic and fitness changes accompanying adaptation of an arbovirus to vertebrate and invertebrate cells. *J. Virol.*, *73*, 4316–4326.

[44] Wilke, C. O. (2001). Adaptive evolution on neutral networks. *Bull. Math. Biol.*, *63*, 715–730.

[45] Wilke, C. O. (2001). Selection for fitness versus selection for robustness in RNA secondary structure folding. *Evolution*, *55*, 2412–2420.

[46] Wilke, C. O., & Adami, C. (2002). The biology of digital organisms. *Trends Ecol. Evol.*, *17*, 528–532.

[47] Wilke, C. O., & Adami, C. (2003). Evolution of mutational robustness. *Mutat. Res.*, *522*, 3–11.

[48] Wilke, C. O., & Ronnewinkel, C. (2001). Dynamic fitness landscapes: Expansions for small mutation rates. *Physica A*, *290*, 475–490.

[49] Wilke, C. O., Ronnewinkel, C., & Martinetz, T. (2001). Dynamic fitness landscapes in molecular evolution. *Phys. Rep.*, *349*, 395–446.

[50] Wilke, C. O., Wang, J. L., Ofria, C., Lenski, R. E., & Adami, C. (2001). Evolution of digital organisms at high mutation rate leads to survival of the flattest. *Nature*, *412*, 331–333.

[51] Zárate, S., & Novella, I. S. (2004). Vesicular stomatitis virus evolution during alternation between persistent infection in insect cells and acute infection in mammalian cells is dominated by the persistence phase. *J. Virol.*, *78*, 12236–12242.

Chapter 4

Tradeoff Between Short-Term and Long-Term Adaptation in a Changing Environment

The work presented here is heavily based on:

R. Forster and C. O. Wilke, Tradeoff between short-term and long-term adaptation in a changing environment. *Physical Review E* 72, 041922 (2005). Reprinted with copyright permissions.

4.1 Abstract

We investigate the competition dynamics of two microbial or viral strains that live in an environment that switches periodically between two states. One of the strains is adapted to the long-term environment, but pays a short-term cost, while the other is adapted to the short-term environment and pays a cost in the long term. We explore the tradeoff between these alternative strategies in extensive numerical simulations and present a simple analytic model that can predict the outcome of these competitions as a function of the mutation rate and the time scale of the environmental changes. Our model is relevant for arboviruses, which alternate between different host species on a regular basis.

4.2 Introduction

The quasispecies model (8) is the premier model to study the evolution of asexual replicators, such as self-replicating molecules or viruses (7, 5). Originally formulated for constant environments, the quasispecies model has recently been extended to describe adaptation to a changing environment (15, 26, 16, 10, 1). The various extensions of the model all work within the original deterministic framework developed by Eigen and co-workers, and thus assume that the population size is infinite. This assumption implies that a population can never lose any genetic information. However, the loss of genetic material and the mechanisms that prevent it from occurring, are probably major forces shaping the evolutionary dynamics of finite populations in time-dependent environments: A finite population can lose to mutation pressure previously useful genetic material that has become meaningless after a change in the environment, and the population may not be able to reacquire this material when the environment changes back to its original state. Consequently, the population will be at a selective disadvantage in comparison to another population that has managed to prevent a similar loss, even if the second population had to pay some short-term cost to keep the useless genetic material. Therefore, in an environment that alternates between two (or more) states, natural selection faces two conflicting agendas—specialization to the current state of the environment, or

adaptation to the long term environment which includes both environmental states.

Here, we study the evolutionary dynamics of competing finite populations of asexual replicators in an environment that alternates between two states, remaining in each state for a time interval of length $T/2$ before switching to the other. To study the tradeoff between the competing forces of selection in this environment, we consider two different strains of replicators, as previously proposed (24). The first strain, which we refer to as the *fused strain*, has a single gene that performs equally well in both environmental states. The second strain, which we refer to as the *divided strain*, has two genes, each of which is advantageous in one environmental state and useless in the other. Clearly, if the fused strain performs as well as the divided strain in both environments, without paying any additional cost, then the divided strain cannot have a selective advantage over the fused strain, regardless of the time scale on which environmental changes happen. If, however, the fused strain does pay some small cost, then which strain is advantageous depends on the exact interplay of the cost, the time scale of environmental change, and the mutation rate. Here, we develop a method to assess and analyze this interplay, and to predict which strain is advantageous in a given setting. As cost, we consider the differential mutation pressure that arises when the fused and the divided strain have genes of different length (27, 23). However, it is straightforward to extend our approach to other types of costs.

Note that while we refer to separate genes throughout this paper, our model can also apply to separate functions carried out by a single gene. In this case, the divided strain corresponds to a gene that can adapt to either function, but not to both at the same time, whereas the fused strain corresponds to a gene that can adapt to both functions at the same time. Such a situation has been observed in an artificial-life simulation with a changing environment (13), where the fusion of genetic function evolved presumably through changes in the amount of epistatic interactions among the different parts of the organisms' genomes.

4.3 Materials and Methods

4.3.1 Model

Here, we model the evolutionary dynamics of a finite population in a time-dependent environment. For comparison, for an *infinite population* in a time-dependent environment, the quasispecies equation reads (26)

$$\frac{dy_i(t)}{dt} = \sum_j w_j(t)\mu_{ij}y_j(t) - y_i(t)\sum_j w_j(t)y_j(t)\,, \tag{4.1}$$

where y_i is the fraction of type i in the population, $w_i(t)$ is the replication rate (i.e., fitness) of type i at time t, and μ_{ij} is the mutation rate per unit time from type j to type i. The quadratic term corresponds to the total production of new organisms per unit time and is subtracted to keep the $y_i(t)$ normalized.

We represent all genes as binary strings. The divided strain has two genes, each of length L_{div}, and the fused strain has a single gene of length L_{fuse}. For each gene, there exists a single functional sequence (the master sequence) that confers the selective advantage, and all alternative sequences are nonfunctional, regardless of the environment. The reproductive fitness $w_i(t)$ of an individual is determined by whether the individual has a functional gene specialized for the current state of the environment. An individual with the correct functional gene has fitness $1 + s$, while an individual without such gene has fitness 1. Mutations occur upon reproduction with a per-site probability μ, corresponding to a per-gene mutation rate of $U = \mu L$.

4.3.2 Simulation

Both the speed of environmental change and the mutation rate μ are important factors in determining the outcome of the competition between the divided and fused strains. To assess their relative importance, we simulated a population of $N = 1000$ individuals reproducing in discrete generations, and with probability of reproduction proportional to their fitness (Wright-Fisher sampling). Initially the population was divided equally between fully functional members of the two strains and the simulation continued until one strain became extinct. We fixed the length of the

divided strain at $L_{\rm div} = 5$, while $L_{\rm fuse}$ varied from 3 to 11. We performed $10,000$ replicates at each pair of period lengths and mutation rates, for $T/2$ $(10, 30, 100, 300, 1000, 3000, 10,000)$ and $U_{\rm div} = \mu L_{\rm div}$ $(0.0001, 0.0003, 0.001, 0.003, 0.01, 0.03, 0.1, 0.3, 1, 3)$.

4.4 Results

4.4.1 Time Scales

To understand the dynamics of competition between the divided and fused strains in a finite population, we consider two time scales—the competition time scale $T_{\rm c}$ and the drift time scale $T_{\rm dr}$. In order to calculate these time scales, we need to know the selective advantage of one strain over the other. The selective advantage s of genotype i over genotype j is a key quantity in theoretical population genetics, and is defined as $s = (w_i - w_j)/w_j$ (4). While this definition is *a priori* applicable only to individual genotypes, it turns out that to a good approximation standard results from population genetics can be applied to separate strains (which consist of a mixture of closely related mutants) if we treat each strain as an individual genotype with fitness given by the strain average $\langle w \rangle$ (22). Throughout this paper, we refer to the selective advantage of one strain over another as the effective selective advantage of this strain. We define the effective fitness advantage of the fused strain over the divided strain by

$$s_{\rm eff} = \frac{\langle w_{\rm fuse} \rangle - \langle w_{\rm div} \rangle}{\min\{\langle w_{\rm fuse} \rangle, \langle w_{\rm div} \rangle\}}. \tag{4.2}$$

This definition guarantees that the magnitude of $s_{\rm eff}$ corresponds to the fitness advantage of the superior strain, while the sign indicates whether the fused strain is superior (positive $s_{\rm eff}$) or inferior (negative $s_{\rm eff}$). If one strain has an effective fitness advantage $|s_{\rm eff}|$ over the other, the competition time scale $T_{\rm c}$ is defined to be the typical time until extinction of the inferior strain in a constant environment. Neglecting finite population effects, and applying eq. (4.1) to strains rather than genotypes, we find that the population fraction $x(t)$ of the superior strain changes approximately

according to the logistic equation

$$\dot{x}(t) = |s_{\text{eff}}|x(t)[1 - x(t)], \tag{4.3}$$

subject to our initial condition that $x(0) = 1/2$. To determine the typical extinction time of the inferior strain, we solve eq. (4.3) for the time when a single member remains of the inferior strain. Thus

$$T_{\text{c}} = \frac{\ln(N-1)}{|s_{\text{eff}}|} \approx \frac{\ln N}{|s_{\text{eff}}|}. \tag{4.4}$$

The drift time scale is defined as the average time for a neutral mutation to go to fixation. Neutral drift becomes important when the fitness advantage between the competing strains is small compared to the fluctuations due to finite sampling effects, i.e., when $s_{\text{eff}} \lesssim 1/N$ (11). For our initial conditions, diffusion theory (12) predicts that

$$T_{\text{dr}} = 2N \ln 2 \approx 1390 \,\text{generations}. \tag{4.5}$$

4.4.2 Quasispecies Effects

Given sufficient time to reach equilibrium, each strain will adopt a quasispecies distribution consisting primarily of those members with a functioning gene adapted to the current environment, together with the deleterious mutants that are constantly regenerated through mutation pressure. In our dynamic fitness landscape, the fused strain will attain this equilibrium distribution after some time, while the divided strain will attempt to equilibrate to the current environment and then go through a period of transition when the environment changes. As an example, figure 4.1 illustrates the formation of the fused-strain quasispecies and the dynamics of the divided-strain quasispecies in the limit of large population size.

Before we can use the time scales derived in the preceeding section to predict the competition's outcome, we must estimate the average fitness of each strain appearing in eq. (4.2). In our model, the average fitness $\langle w \rangle$ of a strain is given by $\langle w \rangle = 1 + sy_0$, where y_0 is the fraction of the population

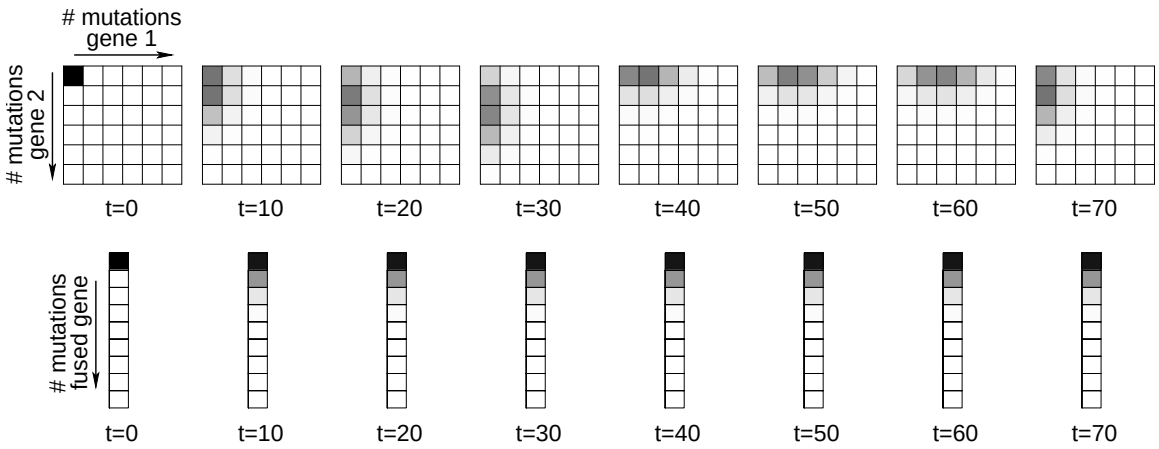


Figure 4.1: Population structure of the divided and fused strains at various time points. Gray levels indicate the fraction of sequences at the given mutational distance from the respective error-free sequence. Parameters are: Oscillation period $T = 60$, per-site mutation rate $\mu = 0.02$, length of a single gene of the divided strain $L_{\text{div}} = 5$, length of the fused gene $L_{\text{fused}} = 8$, selective advantage of functional gene $s = 1$, infinite population size. From (24).

that has a functional gene for the current environment. Thus eq. (4.2) becomes

$$s_{\text{eff}} = \frac{sy_{0,\text{fuse}} - sy_{0,\text{div}}}{1 + s\min\{y_{0,\text{fuse}}, y_{0,\text{div}}\}}\,. \tag{4.6}$$

To estimate y_0, we assume that the quasispecies immediately reaches its equilibrium distribution. Denoting $y_i(t)$ as the population fraction of a given strain with i errors at time t, we neglect the back mutation term (that is, the term that represents mutations from genotypes with errors to error-free genotypes) in eq. (4.1) and obtain

$$\dot{y}_0(t) = (1+s)Q_{00}(L)y_0(t) - [1 + sy_0(t)]y_0(t), \tag{4.7}$$

where $Q_{ij}(L)$ is the probability that a string of length L with j errors mutates into one with i errors. $Q_{ij}(L)$ has been given for example in (28). Setting $\dot{y}_0 = 0$ in equilibrium, we find $y_0 = [(1+s)Q_{00}(L) - 1]/s$. However, when back mutations become significant and y_0 approaches zero in this expression, we reach the classical error threshold (8). For mutation rates beyond this point, we

assume that the population is randomized uniformly over all possible states, and hence we use

$$y_0 = \max\left\{\frac{(1+s)Q_{00}(L)-1}{s}, 2^{-L}\right\} \tag{4.8}$$

for all mutation rates. For mutation rates below the error threshold, eq. (4.8) yields a simple form for the magnitude of the fitness advantage, $|s_{\text{eff}}| = (1-\mu)^{-|L_{\text{div}}-L_{\text{fuse}}|} - 1$, while the sign of s_{eff} is given by $\text{sgn}(L_{\text{div}} - L_{\text{fuse}})$. This result shows that the effect of mutational load on fitness is to favor whichever strain has the shorter length.

The result of Eq (4.8) applies to a quasispecies that has reached equilibrium. While the equilibrium assumption provides a good estimate for the fused strain, the rate of environmental changes may prevent the divided strain's quasispecies from ever reaching this equilibrium. If the environment changes quickly (relative to the competition and drift time scales), the divided strain persists in an average environment that requires both genes for functionality (26), and hence this strain has an effective gene length of $2L_{\text{div}}$. In this case, we approximate the resulting quasispecies as one with a single gene of length $2L_{\text{div}}$ and replace $Q_{00}(L_{\text{div}})$ by $Q_{00}(2L_{\text{div}})$ in eq. (4.8). Even though this approximation disregards the two-dimensional nature of the divided-strain quasispecies, it gives a reasonably good estimate of the true y_0 for most mutation rates. Note, however, that we do not replace $2^{-L\text{div}}$ by $2^{-2L\text{div}}$. At any point in time, the environment favors only one of the divided strain's two genes, and hence beyond the error threshold the probability that a randomly chosen individual from the divided strain carries a functional gene remains $2^{-L\text{div}}$. We refer to s_{eff} calculated with $Q_{00}(L_{\text{div}})$ as the short-term limit, and to s_{eff} calculated with $Q_{00}(2L_{\text{div}})$ as the long-term limit.

4.4.3 Predicting the Probability of Fixation

We propose a simple ternary model to predict the probability of fixation p of the fused strain. In our model, p is 0 if the divided strain is favored, 1/2 for neutral evolution, when both strains are equally likely to go to fixation, or 1 if the fused strain is favored. First, we classify the selective

Table 4.1: Selective regime, as determined by the relative magnitudes of $T_{\rm dr}$, $T_{\rm c}$, and $T/2$

Condition	Selective Regime
$T/2 > T_{\rm c}$	short-term limit
$T_{\rm dr} < T/2 < T_{\rm c}$	neutral limit
$T/2 < T_{\rm dr}\,, T_{\rm c}$	long-term limit

Table 4.2: Model predictions, as determined by the relative magnitude of $s_{\rm eff}$ and $1/N$

Fitness Advantage (Fused Strain)	Prediction	p
$s_{\rm eff} \leq -1/N$	divided strain wins	0
$-1/N < s_{\rm eff} < 1/N$	neutral*	1/2
$s_{\rm eff} \geq 1/N$	fused strain wins	1

* In the neutral case, both strains are equally likely to win.

regime based on the drift time $T_{\rm dr}$ and the short-term competitive time scale $T_{\rm c} = \ln N/|s_{\rm eff}^{\rm short}|$ in comparison to $T/2$, the length of time for which the environment remains constant (see table 4.1). If $T_{\rm c} < T/2$, then we expect the competition between the two strains to end before the environment changes even once and hence the short-term limit applies. The value of $T_{\rm dr}$ is irrelevant in this case. If both times are longer than $T/2$, then we expect the competition to extend over several half periods and hence the long-term limit applies. Finally, if $T/2$ is smaller than $T_{\rm c}$, but larger than $T_{\rm dr}$, then we expect drift to be the dominant force. We call this regime the neutral limit and set $s_{\rm eff} = 0$. Having determined the appropriate limiting case (short-term, neutral, or long-term limit), we can use the associated fitness advantage $s_{\rm eff}$ to predict the probability p of fixation for the fused strain (table 4.2). If $|s_{\rm eff}| < 1/N$, then the two strains are effectively neutral. Hence, the outcome of the competition is determined by drift and $p = 1/2$. Otherwise, $p = 0$ or 1 depending on whether $s_{\rm eff}$ is negative or positive.

4.4.4 Comparison with Simulation Results

From our simulations, we estimated the probability p that the fused strain would fixate in the population as a function of $T/2$, $U_{\rm div}$, and $L_{\rm fuse}$. Results are shown in figure 4.2A,B,C, for representative values of $L_{\rm fuse}$. The probabilities obtained by simulation have a standard error of $\pm 1\%$. Figure 4.2D,E,F shows the corresponding predictions of our model. Over the full range of parame-

ters described in section 4.3.2, our model predicted that the divided strain was superior in 235 cases ($p = 0$), that the strains were neutral in 240 cases ($p = 1/2$), and that the fused strain was superior in 155 cases ($p = 1$). The range of simulation results corresponding to each of these three predictions is shown in figure 4.3. When the model predicted $p = 0$, 85% of all simulation probabilities fell in the range 0–0.1. When $p = 1$ was predicted, 84% of simulation probabilities fell between 0.9 and 1, while in the neutral case ($p = 1/2$ predicted), 58% of simulation probabilities fell between 0.45 and 0.55. The rms error between the simulation values and the model predictions is 18.6%, averaged over all cases. For comparison, the best possible rms error for any such ternary model is 7.6% on this data.

4.5 Discussion

The study of quasispecies dynamics in a time dependent fitness landscape to date has primarily focused on the limit of infinite population size (15, 26, 16, 10, 1). In a periodic fitness landscape, an infinite population size guarantees that competition between two strains will result in the deterministic extinction of the inferior strain or, in certain finely tuned cases, an unstable coexistence between the strains (although frequency dependent selection may stabilize this equilibrium (25)). In contrast, the generalization of these models to a finite population presents a continuous range of possibilities from almost certain extinction to the complete randomness of neutral drift.

In this study, we present a model for predicting the outcome of competition between finite quasispecies' in a periodic environment. As applied to our specific case of competition between a divided and a fused strain, our model shows good qualitative and quantitative results in comparison with simulation (figs. 4.2 and 4.3). When $L_{\text{fuse}} < L_{\text{div}}$ or $L_{\text{fuse}} > 2L_{\text{div}}$, one of the two strains is strictly superior and the outcome of the competition is determined by whether drift or competition are more important at the given mutation rate. For intermediate values of L_{fuse} the competitive dynamics are more complex, as certain combinations of mutation rate and period lengths favor the fused strain while others favor the divided strain.

The only qualitative feature of the competitive dynamics we have ignored is the time for the

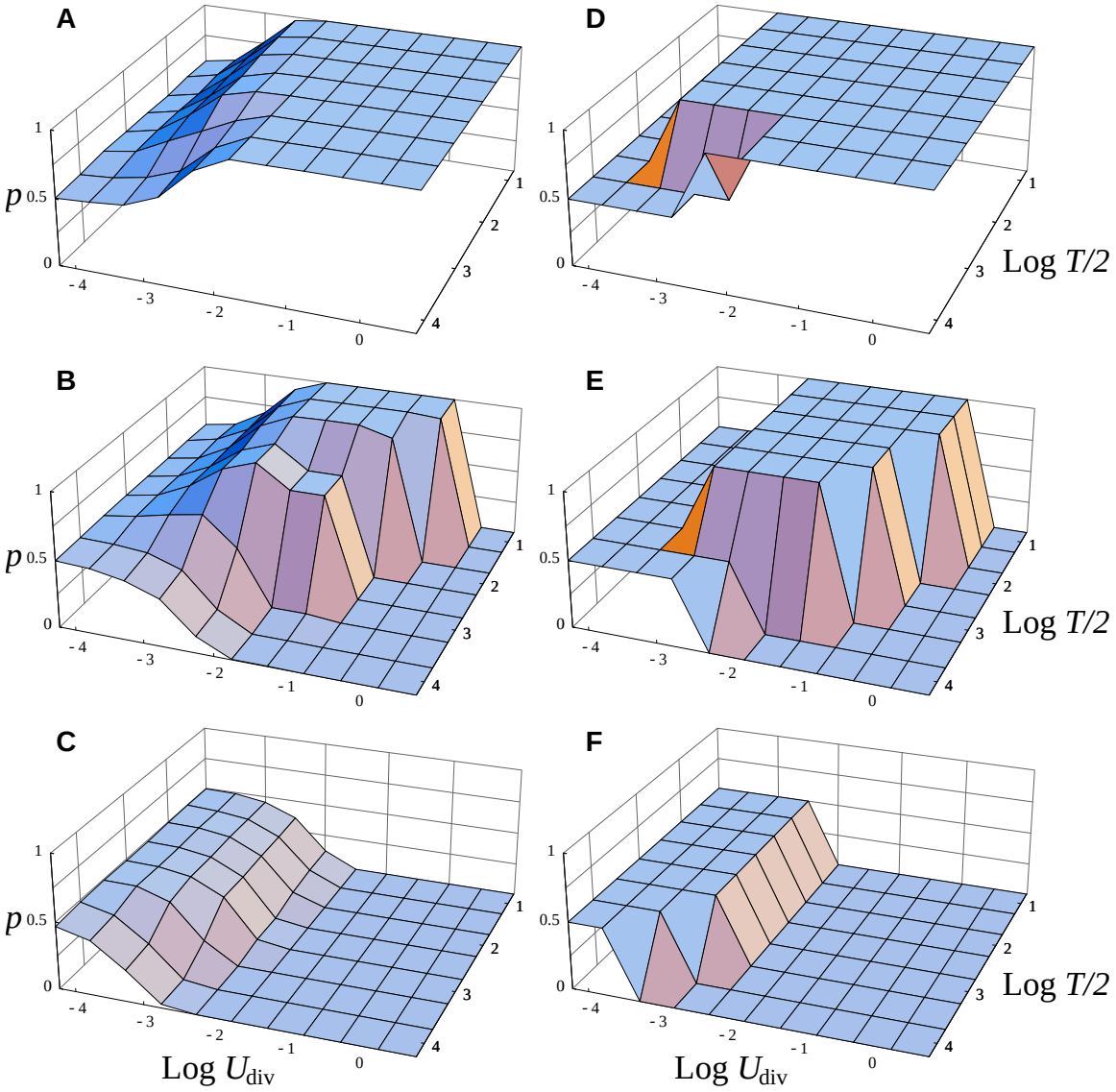


Figure 4.2: Left (A, B, C): Simulation results for the probability of fixation of the fused strain as a function of the mutation rate $U_{\text{div}} = \mu L_{\text{div}}$ and period length T for $L_{\text{div}} = 5$ and $L_{\text{fuse}} = 4, 6, 11$ (top to bottom). Simulation results have a standard error of approximately $\pm 1\%$. Right (D, E, F): Model predictions for the same. Parameter values are $s = 1$, $N = 1000$.

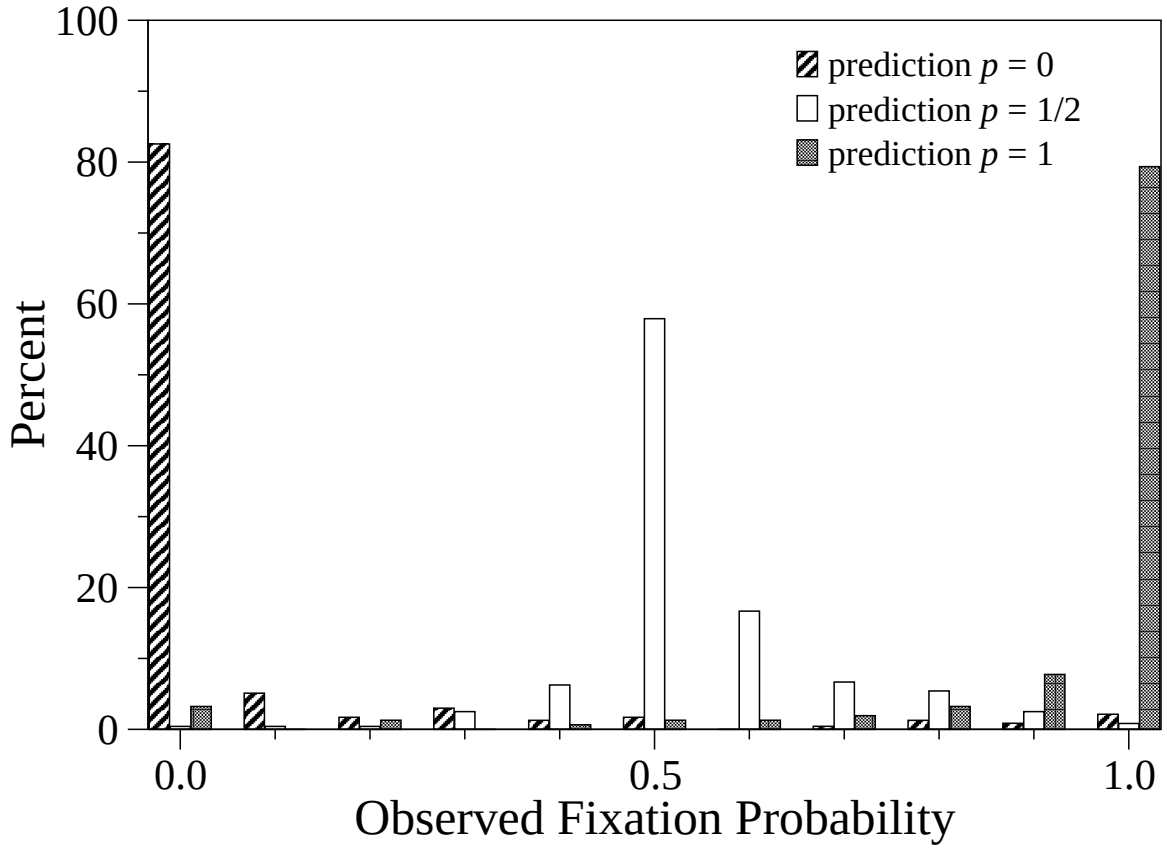


Figure 4.3: Fixation probabilities for the fused strain (as determined by simulation), classified by model prediction. The model predicted $p = 0$ in 235 cases, $p = 1/2$ in 240 cases, and $p = 1$ in 155 cases. The x axis is binned in 0.1 increments, with the bins' midpoint value shown.

quasispecies to reach its equilibrium distribution. We decided not to include a quasispecies time scale in our model after preliminary efforts showed that the added complexity failed to significantly improve model accuracy. While the fused strain quickly reaches its equilibrium independent of the environment's state, the divided strain may fare worse than predicted in some special cases when it spends significant time in transitions between the quasispecies distributions for each environment, being poorly adapted in the meanwhile. Still, these cases are relatively rare since at low mutation rates quasispecies effects are less important, while at high mutation rates the error threshold masks any such effects.

Our results suggest that, under appropriate circumstances, a selective pressure exists to fuse or divide complementary genes in a periodic environment. The tendency of uncertain enviromental conditions to facilitate large scale genetic changes such as this one has recently been studied (6). Most genomes are full of apparently useless or nonfunctional genetic material, which in our model corresponds to the excess length of the fused strain over a single gene of the divided strain. In populations as small as 1000 such as those we studied, such temporarily useless genetic material comprising 20% of the genome could be stably maintained for periods of dormancy up to 1000 generations (figure 4.2B). Alternatively, a fused gene exposed to an appropriate periodic environment might undergo division in response to the selective pressure we describe. The beginnings of a genome segmentation has been observed in foot-and-mouth disease virus (FMDV) in response to conditions of high multiplicity of infection (9). One explanation of this segmentation is that translational speed favors shorter genes (14, 19), a fitness effect similar to the length-based mutational pressure of our model.

Our model applies directly to the evolution of arboviruses, which are viruses transmitted by arthropods. For example, the arbovirus West Nile virus is transmitted from birds to birds (and the occasional human) by mosquitoes, experiencing the alternating environments of the avian and insect hosts. Experimental virologists have long tried to determine whether viruses subjected to such alternating environments adapt to the short-term or the long-term environment, but have not found a conclusive answer. Experiments with vesicular stomatitis virus (VSV), eastern equine encephalitis

virus (EEEV), and Dengue virus in cells of insect and mammalian origin have shown that in some cases, adaptation to one cell type leads to loss of fitness in the other cell type, while in other cases fitness can increase in both cell types at the same time (17, 18, 21, 20, 3, 2, 29). Which of the two cases occurs depends on the time spent in each of the two hosts, and also on the details of the fitness landscape in the two hosts. It stands to reason that in future experiments in which the time scale of environmental change is varied over a wide range, a switch in the adaptation strategy from short term to long term will be observed, and that the time scale at which the switch occurs can be predicted with the methods we have developed here.

In conclusion, we have demonstrated that the competitive dynamics of finite populations in a time-dependent environment can be quite complex, but that nevertheless, estimates of the effective fitness advantages of the different strains together with an understanding of the drift and competitive time scales can lead to remarkably accurate predictions of the evolutionary dynamics. We believe that similar techniques will prove useful to interpret and predict outcomes of virus-evolution experiments in changing environments.

Bibliography

[1] Brumer, Y., & Shakhnovich, E. I. (2004). Host-parasite coevolution and optimal mutation rates for semiconservative quasispecies. *Phys. Rev. E*, *69*, 061909.

[2] Chen, W.-J., Wu, H.-R., & Chiou, S.-S. (2003). E/NS1 modifications of dengue 2 virus after serial passages in mammalian and/or mosquito cells. *Intervirology*, *46*, 289–295.

[3] Cooper, L. A., & Scott, T. W. (2001). Differential evolution of eastern equine encephalitis virus populations in response to host cell type. *Genetics*, *157*, 1403–1412.

[4] Crow, J. F., & Kimura, M. (1970). *An Introduction to Population Genetics Theory*. New York: Harper & Row.

[5] Domingo, E., Biebricher, C. K., Eigen, M., & Holland, J. J. (2001). *Quasispecies and RNA Virus Evolution: Principles and Consequences*. Georgetown: Landes Bioscience.

[6] Earl, D. J., & Deem, M. W. (2004). Evolvability is a selectable trait. *Proc. Natl. Acad. Sci. USA*, *101*, 11531–11536.

[7] Eigen, M., McCaskill, J., & Schuster, P. (1988). Molecular quasi-species. *J. Phys. Chem.*, *92*, 6881–6891.

[8] Eigen, M., & Schuster, P. (1979). *The Hypercycle—A Principle of Natural Self-Organization*. Berlin: Springer-Verlag.

[9] García-Arriaza, J., Manrubia, S. C., Toja, M., Domingo, E., & Escarmís, C. (2004). Evolutionary transition toward defective RNAs that are infectious by complementation. *J. Virol.*, *78*, 11678–11685.

[10] Kamp, C., & Bornholdt, S. (2002). Coevolution of quasispecies: B-cell mutation rates maximize viral error catastrophes. *Phys. Rev. Lett.*, *88*, 068104.

[11] Kimura, M. (1964). Diffusion models in population genetics. *J. Appl. Prob.*, *1*, 177–232.

[12] Kimura, M., & Ohta, T. (1969). The average number of generations until fixation of a mutant gene in a finite population. *Genetics*, *61*, 763–771.

[13] Li, Y., & Wilke, C. O. (2004). Digital evolution in time-dependent fitness landscapes. *Artificial Life*, *10*, 123–134.

[14] Nee, S. (1987). The evolution of multicompartmental genomes in viruses. *J. Mol. Evol.*, *25*, 277–281.

[15] Nilsson, M., & Snoad, N. (2000). Error thresholds on dynamic fitness landscapes. *Phys. Rev. Lett.*, *84*, 191–194.

[16] Nilsson, M., & Snoad, N. (2002). Quasispecies evolution on a fitness landscape with a fluctuating peak. *Phys. Rev. E*, *65*, 031901.

[17] Novella, I. S., Clarke, D. K., Quer, J., Duarte, E. A., Lee, C. H., Weaver, S. C., Elena, S. F., Moya, A., Domingo, E., & Holland, J. J. (1995). Extreme fitness differences in mammalian and insect hosts after continuous replication of vesicular stomatitis virus in sandfly cells. *J. Virol.*, *69*, 6805–6809.

[18] Novella, I. S., Hershey, C. L., Escarmis, C., Domingo, E., & Holland, J. (1999). Lack of evolutionary stasis during alternating replication of an arbovirus in insect and mammalian cells. *J. Mol. Biol.*, *287*, 459–465.

[19] Szathmary, E. (1992). Natural selection and dynamical coexistence of defective and complementing virus segments. *J. Theor. Bio.*, *157*, 383–406.

[20] Turner, P. E., & Elena, S. F. (2000). Cost of host radiation in an RNA virus. *Genetics*, *156*, 1465–1470.

[21] Weaver, S. C., Brault, A. C., Kang, W., & Holland, J. J. (1999). Genetic and fitness changes accompanying adaptation of an arbovirus to vertebrate and invertebrate cells. *J. Virol.*, *73*, 4316–4326.

[22] Wilke, C. O. (2001). Selection for fitness versus selection for robustness in RNA secondary structure folding. *Evolution*, *55*, 2412–2420.

[23] Wilke, C. O., & Adami, C. (2003). Evolution of mutational robustness. *Mutat. Res.*, *522*, 3–11.

[24] Wilke, C. O., Forster, R., & Novella, I. S. (2006). Quasispecies in time-dependent environments. *Curr. Top. Microbiol. Immunol.*, *299*, 33–50.

[25] Wilke, C. O., Ressig, D. D., & Novella, I. S. (2004). Replication at periodically changing multiplicity of infection promotes stable coexistence of competing viral populations. *Evolution*, *58*, 900–905.

[26] Wilke, C. O., Ronnewinkel, C., & Martinetz, T. (2001). Dynamic fitness landscapes in molecular evolution. *Phys. Rep.*, *349*, 395–446.

[27] Wilke, C. O., Wang, J. L., Ofria, C., Lenski, R. E., & Adami, C. (2001). Evolution of digital organisms at high mutation rate leads to survival of the flattest. *Nature*, *412*, 331–333.

[28] Woodcock, G., & Higgs, P. G. (1996). Population evolution on a multiplicative single-peak fitness landscape. *J. Theor. Biol.*, *179*, 61–73.

[29] Zárate, S., & Novella, I. S. (2004). Vesicular stomatitis virus evolution during alternation between persistent infection in insect cells and acute infection in mammalian cells is dominated by the persistence phase. *J. Virol.*, *78*, 12236–12242.

Chapter 5

Frequency-Dependent Selection in a Periodic Environment

Submitted to *Physical Review E*, March 2006.

Authors as published: Robert Forster and Claus O. Wilke.

5.1 Abstract

We examine the action of natural selection in a periodically changing environment where two competing strains are specialists respectively for each environmental state. When the relative fitness of the strains is subject to a very general class of frequency-dependent selection, we show that coexistence rather than extinction is the likely outcome. This coexistence may be a stable periodic equilibrium, stable limit cycles of varying lengths, or be deterministically chaotic. Our model is applicable to the population dynamics commonly found in many types of viruses.

5.2 Introduction

Alternating environmental conditions are common in many viral systems. Arboviruses are a class of RNA viruses that alternate host species between arthropods and vertebrates, and include examples such as West Nile virus, dengue fever virus, and vesicular stomatitis virus (VSV). Experimental efforts have demonstrated that adaptation and evolution occur regularly under alternating environmental conditions (3, 4, 35, 21, 22, 36, 39), suggesting that the wild type and mutant strains present are in direct competition. Similarly, recent studies have found that density-dependent selection is present in many viral experiments (32, 23, 37), and the alternating stages of low and high multiplicities of infection (m.o.i.) can also present a periodic environment to the viral population (37).

Also common in many viral systems is the presence of frequency-dependent selection (9, 38, 34). While the frequency dependence of a strain's fitness may take many functional forms, a common type is where the strain competes best when it is rare relative to its competitor. This form of "fittest when rare" frequency dependence naturally arises under biological conditions of parasitism or through the evolution of cheating strategies. Viral deletion mutants known as defective interfering particles (DIPs) are viral parasites commonly found under conditions of high m.o.i. (14, 25, 27, 11). No longer capable of independent reproduction themselves, DIPs rely on co-opting wild-type viral products in an infected cell, and have a selective advantage over the wild type due to the faster replication

of their smaller genome (19, 28). Game theoretic models have been applied to conditions of viral coinfection, where selfish behavior similar to that in the prisoner's dilemma evolved (33, 17, 34). Such selfish strategies naturally work best rare, suffering when common from a lack of others to exploit.

Motivated by these common features of viral systems, we develop a theoretical model to study competition between two strains in a periodic environment. Each strain is relatively well adapted to one environment, but poorly adapted to the other. In this context, we model the general effects of frequency-dependent selection under the simplifying assumption that the strains compete best when rare. For a wide range of competitive conditions, our model predicts coexistence of both strains. This coexistence may be as simple as a periodic alternation of population sizes synchronized with environmental changes, or as complex as fully chaotic population dynamics.

5.3 Model

The competition between two strains, A and B, occurs in a time-dependent environment which oscillates between two distinct states. We assume the limit of large population size and describe the current state of the population by x, the population fraction of strain A. To characterize the fitness of each strain, we normalize the fitness of the superior strain to be 1, and describe the relative fitness of the inferior strain by a frequency-dependent fitness function $w(x)$. During environmental state 1, therefore, strain A is superior and has a constant normalized fitness $w_A = 1$, while strain B is inferior with frequency-dependent fitness $0 < w_B(x) < 1$. Consistent with our assumption that strain B competes best when rare, we take $w_B(x)$ to be a monotonically increasing function. For definiteness, we denote relative fitness of strain B when very common as $b_0 = w_B(0)$ and the fitness of strain B when very rare as $b_1 = w_B(1)$, where $b_0 \leq b_1$ by assumption. If the population starts in environment 1 with a population fraction of strain A given by x, it will have a fraction $f_1(x)$ at the end given by

$$f_1(x) = \frac{x}{x + (1-x)w_B(x)} \,. \tag{5.1}$$

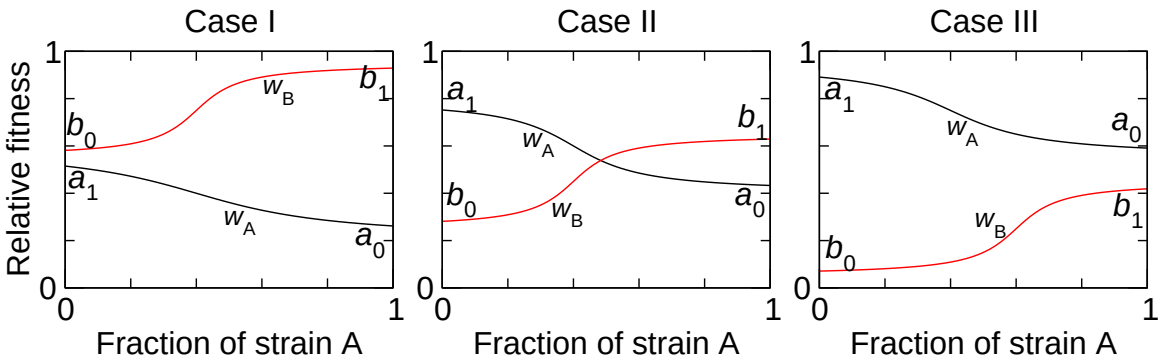


Figure 5.1: Three cases of qualitatively different relationships between the frequency dependence of fitness functions $w_A(x)$ and $w_B(x)$. Case I, left: Strain B is strictly superior. Case II, middle: Neither strain is strictly superior. Case III, right: Strain A is strictly superior.

In environmental state 2, the roles of the strains reverse. Strain B is superior with normalized fitness $w_B = 1$ while strain A is inferior with frequency-dependent fitness $w_A(x)$. This means that $w_A(x)$ is a monotonically decreasing function, and we denote the values $a_1 = w_A(0)$ and $a_0 = w_A(1)$ where $a_0 \leq a_1$. Let $f_2(x)$ give the population fraction at the close of environment 2, given that it began environment 2 with population x. Then we have

$$f_2(x) = \frac{x\, w_A(x)}{(1-x) + x w_A(x)}\,. \tag{5.2}$$

We will make the biologically reasonable assumption that both fitness functions $w_A(x)$ and $w_B(x)$ are continuous. Under these assumptions, there are three qualitatively different relationships possible between the relative fitness functions w_A and w_B: One or the other can be strictly larger for all values of x, or the fitness functions can cross at some intermediate value of x. These three cases are show schematically in fig. 5.1. Note that the applicable case can be determined solely by considering the relative magnitudes of the four constants a_0, a_1, b_0, and b_1.

5.4 Results

5.4.1 Existence of Equilibria

We prove that the cases I and III from fig. 5.1 allow for no stable coexistence between strains, and extinction occurs as the population fraction converges to $x = 0$ or $x = 1$ respectively. Conversely, in case II there is always an equilibrium value of $x \in (0, 1)$ where both strains coexist. The stability of this equilibrium, however, is more complicated and will be addressed later. If we assume the population begins a period at x in environment 1, it will end the period with population $g(x)$ given by

$$g(x) = f_2[f_1(x)] = \frac{x\, w_A(y)}{w_B(x) + x[w_A(y) - w_B(x)]}\,, \tag{5.3}$$

where $y = f_1(x)$. An equilibrium in the periodic environment corresponds to a solution of the equation $g(x) = x$. We shall henceforth refer to such a solution as an equilibrium, even though the population is actually alternating between two different values, x and $f_1(x)$, as the environmental state changes. We shall refer to a *periodic equilibrium* as a situation where, for example, $g(x)$ alternates between two different values (and the population alternates between four different values before repeating the pattern). An equilibrium point x will be a locally stable or attracting equilibrium if $|g'(x)| < 1$ and locally unstable or repelling if $|g'(x)| > 1$. Assuming the equilibrium value does not correspond to extinction, $x \neq 0, 1$, we find that there can be a solution to $g(x) = x$ if and only if

$$w_A(y) = w_B(x)\,. \tag{5.4}$$

In the cases I or III, the ranges of $w_A(x)$ and $w_B(x)$ are disjoint, and hence there can be no solution to eq. (5.4). Therefore the only equilibrium solutions in these cases are $x = 0$ or 1, and testing the derivative $g'(x)$ at these values confirms that $x = 0$ is attracting while $x = 1$ is repelling in case I, and vice versa for case III. Qualitatively, strain B wins the competition in case I and strain A wins in case III. Observe that this results corresponds to the strain with the strictly larger relative fitness function in fig. 5.1 going to fixation in the population.

To see that there must be an equilibrium in case II, consider the function $h(x) = w_A[f_1(x)] - w_B(x)$. This function is continuous, with $h(0) = a_1 - b_0 > 0$ and $h(1) = a_0 - b_1 < 0$. Thus by the Intermediate Value Theorem, there exists an $x_e \in (0, 1)$ where $h(x_e) = 0$. By construction, this x_e satisfies eq. (5.4) and hence x_e is an equilibrium corresponding to coexistence of both strains. The values $x = 0$ and $x = 1$, corresponding to extinction of either strain, are repelling equilibria. This result can be seen by evaluating $g'(x)$, where we find $g'(0) = b_1/a_0 > 1$ and $g'(1) = a_1/b_0 > 1$. Thus in case II, neither strain will suffer extinction in an infinite population. However, the equilibrium at x_e need not be stable in this case. For example, there could be an attracting periodic cycle or other more complex behavior. In the case of many simple fitness functions we can rule out these complex behaviors, as we show in the next section.

The stable coexistence of strains in case II is more remarkable given how likely it is to occur. While the precise relative probabilities of the three cases will depend on the specific biological system in question, we can qualitatively address this issue by assuming the endpoints of the fitness functions a_0, a_1, b_0, and b_1 are chosen uniformly at random in $[0, 1]$, consistent with the requirement that $a_0 \leq a_1$ and $b_0 \leq b_1$. Under these assumptions, basic probabilistic considerations show that we expect case II to occur 2/3 of the time, while cases I and III occur with probability 1/6 each.

5.4.2 Stable Coexistence in the Linear Case

Consider the special case when the functions $w_A(x)$ and $w_B(x)$ are linear functions,

$$w_A(x) = a_1 - (a_1 - a_0)x\,, \tag{5.5}$$

$$w_B(x) = b_0 + (b_1 - b_0)x\,. \tag{5.6}$$

In this case, we shall prove that there is a unique stable equilibrium when the fitness functions meet the conditions of case II. To see that the equilibrium is unique, consider the function $w_A(y)$

that appears in eq. (5.4). Computing its derivative, we find that that it is a decreasing function,

$$\frac{dw_A(y)}{dx} = \frac{dw_A(y)}{dy}\frac{df_1(x)}{dx} < 0\,. \tag{5.7}$$

While $dw_A(y)/dy$ is always negative since $w_A(y)$ is a decreasing function, the result of eq. (5.7) only holds because $df_1(x)/dx$ is strictly positive in the linear case. Therefore, when we solve the equilibrium equation $w_A(y) = w_B(x)$, we are seeking the intersections between a strictly increasing function, $w_B(x)$, and a strictly decreasing function, $w_A(y)$. Assuming we are in case II where an equilibrium exists for $x \in (0, 1)$, the equilibrium must be unique.

Moreover, this equilibrium is stable. Observe that for any fitness function, including nonlinear ones, $f_2(x)$ is increasing since $df_2(x)/dx > 0$. In the linear case, direct calculation shows that $df_1(x)/dx > 0$ and therefore $g(x)$ is also an increasing function,

$$\frac{dg}{dx} = f_2'[f_1(x)]\frac{df_1(x)}{dx} > 0\,. \tag{5.8}$$

It is impossible for the increasing function $g(x)$ to have any oscillatory behavior. Since the values $x = 0$ and $x = 1$ are known to be repelling fixed points for case II, the remaining equilibrium must be a stable fixed point.

5.4.3 A Stability Condition

Our above proof that a unique and stable equilibrium exists in the case of linear fitness functions can be generalized. The key condition is proving that $g(x)$ is increasing. While $df_2(x)/dx > 0$ holds for all fitness functions, $f_1(x)$ may not be increasing. In general, we have

$$\frac{df_1(x)}{dx} = \frac{w_B(x) - x(1-x)w_B'(x)}{[x + (1-x)w_B(x)]^2}\,. \tag{5.9}$$

Although $w_B'(x) > 0$ by assumption, $f_1'(x)$ could be negative if the fitness function $w_B(x)$ changes very rapidly as a function of the population fraction x. A sufficient but not necessary condition for

stable coexistence is therefore

$$w_B(x) \geq x(1-x)w'_B(x) \quad \forall x \in [0,1]\,. \tag{5.10}$$

We now give an example where rapid change in the frequency-dependent fitness leads to more complex population dynamics.

5.4.4 Chaotic Dynamics in the General Case

In general, the population dynamics of our system can be chaotic. As an example, consider the case that strain A has no frequency dependence, that is, $w_A(x)$ is constant. We take $w_B(x)$ to smoothly change from a straight line to a step function as the parameter a increases. The fitness functions for this example are shown in fig. 5.2, including $w_B(x)$ for varying values of the parameter a. The specific functions used in this example are

$$w_A(x) = \frac{2}{5}, \quad w_B(x) = \frac{1}{2} + \frac{4}{5\pi}\arctan\Big[a\Big(x - \frac{1}{2}\Big)\Big]\,. \tag{5.11}$$

An important result in the theory of one-dimensional iterated maps is Sarkovskii's Theorem (29), sometimes referred to as "Period Three Implies Chaos" (18). This theorem states that whenever a continuous map has a point of period 3 it has points of all period lengths, a common property of chaotic systems. Points of period 3 appear in our example once the parameter $a \gtrsim 10^2$ (see also fig. 5.3). For instance, when $a = 10^3$, we are in the domain of an attracting 8 cycle. In this case, there are still points of period 3 (and hence all periods), but these other periodic points are unstable. One of these unstable period 3 cycles is shown in fig. 5.4 converging quickly to the stable 8 cycle due to finite precision effects. Several other population trajectories are shown in fig. 5.4, illustrating the population dynamics for different parameter values a.

In fig. 5.5, we display the orbits of representative initial points after a large number of iterations (see appendix C for technical details). As the parameter a increases, the attracting equilibrium of the system is no longer a fixed point, but instead becomes a periodic equilibrium. The bifurcations

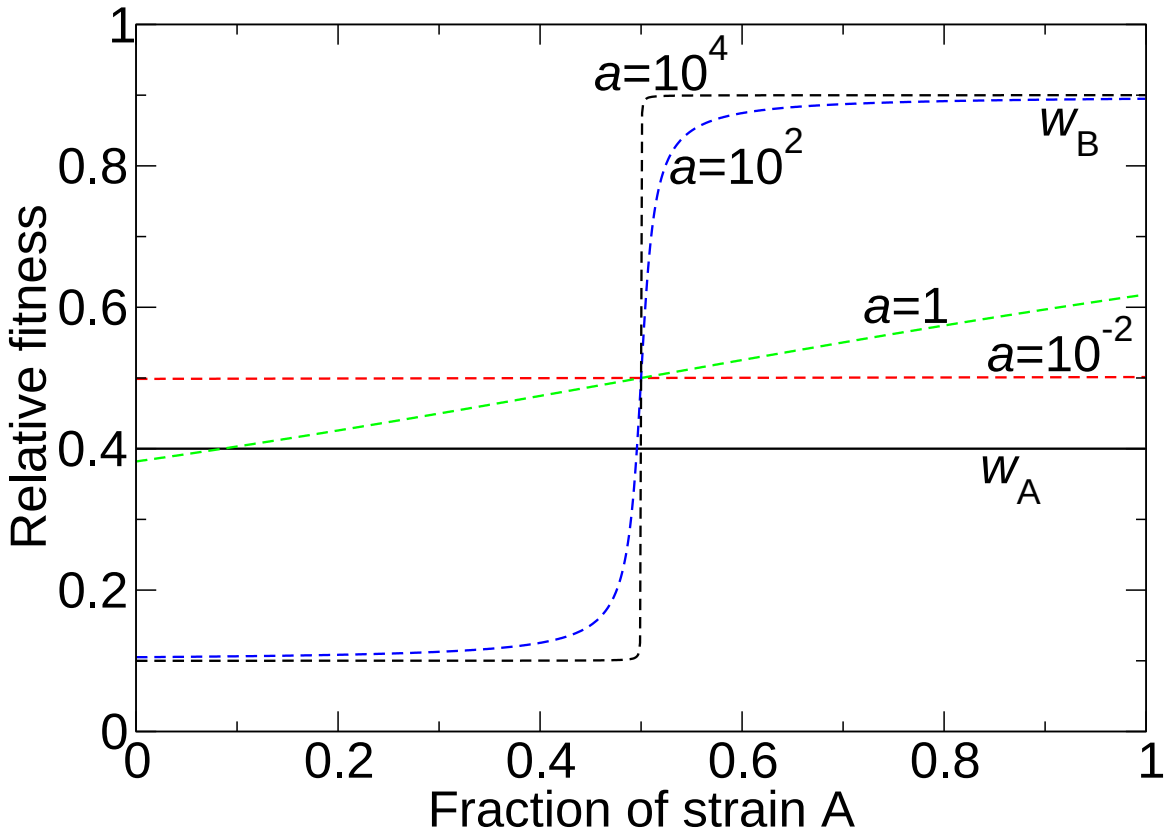


Figure 5.2: Sample fitness function $w_A(x)$ (solid line) and functions $w_B(x)$ (broken lines) given in eq. (5.11). Parameter values for $w_B(x)$ are $a = 10^{-2}, 1, 10^2, 10^4$.

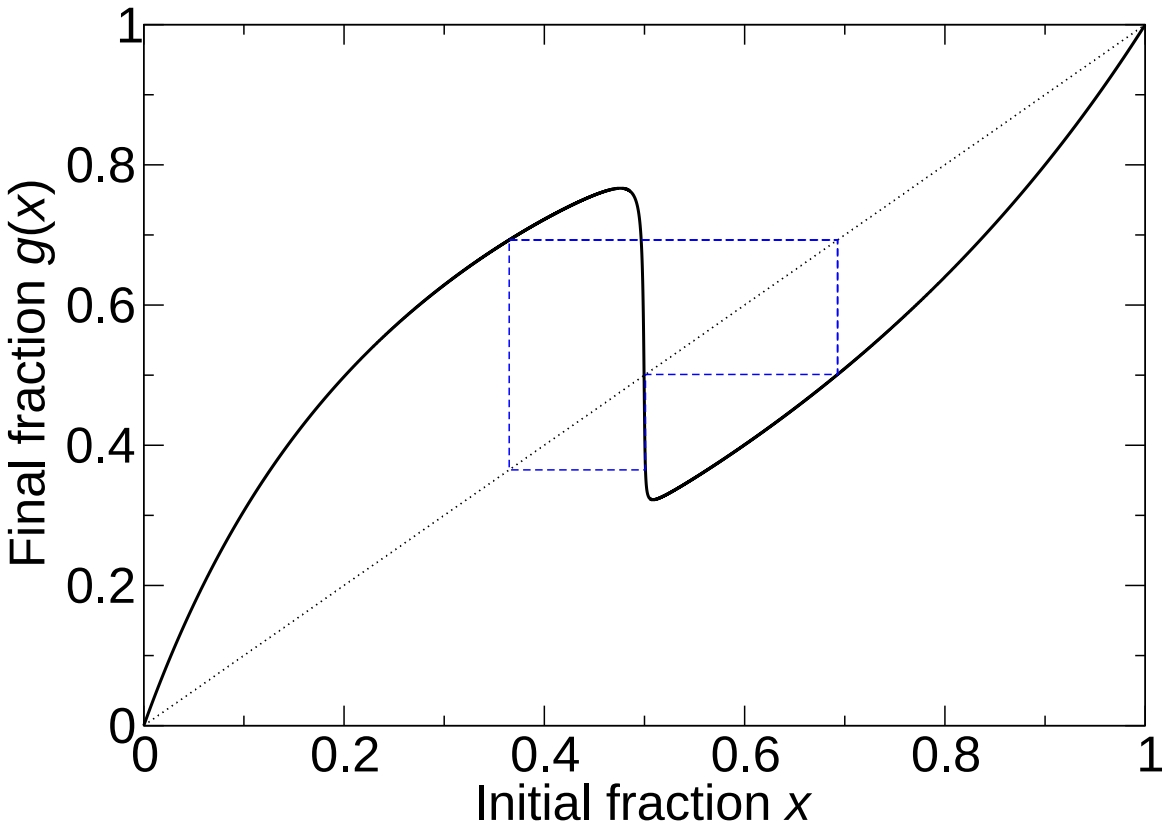


Figure 5.3: Illustration of a 3 cycle on the plot of $g(x)$, using the fitness functions in eq. (5.11) and the parameter value $a = 10^3$. The existence of a 3 cycle often indicates the presence of chaos in a system.

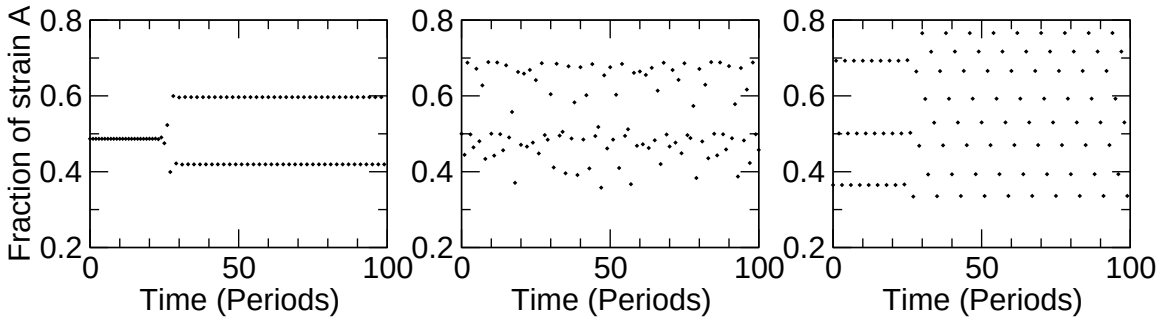


Figure 5.4: Population fraction of strain A, as a function of the number of environmental periods. Left to Right: The unstable equilibrium is attracted to the stable 2 cycle ($a = 10^{1.5}$); the orbit of $x = 1/2$ shows unpredictable behavior ($a = 10^2$); an unstable 3 cycle is attracted to the stable 8 cycle ($a = 10^3$).

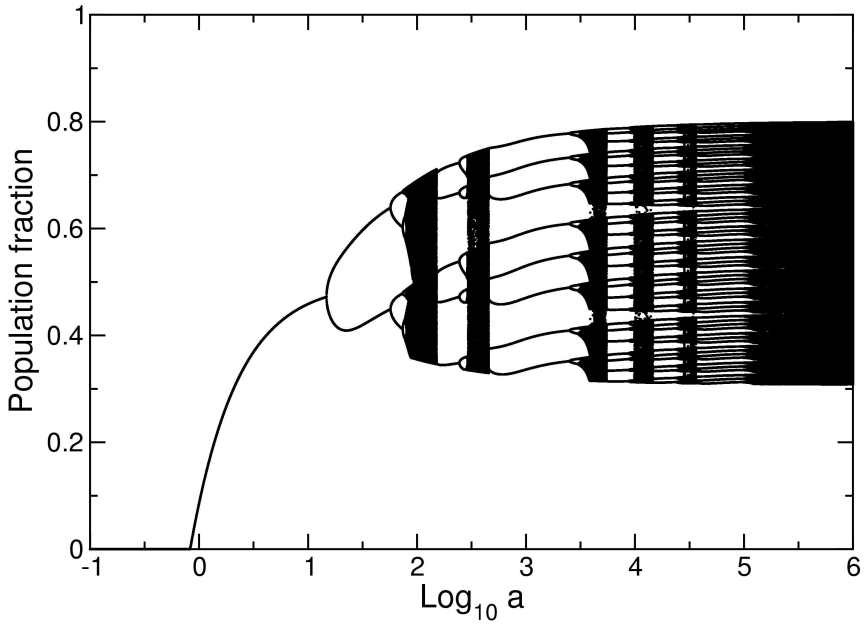


Figure 5.5: Population fraction in steady state, plotted as a function of the parameter a in eq. (5.11). We plot the population fraction x for many periods, after allowing a long time for initial equilibration (periods shown are 1000-2000). From left to right, the system shows extinction of one strain, stable coexistence, periodic behavior, and finally chaos. Initial conditions are chosen so that all attracting orbits will be found (see appendix C).

that occur in fig. 5.5 with increasing a are one of the classic hallmarks of a chaotic system (6). For small values of $a \lesssim 1$, the crossing condition of case II is not met, $x = 0$ is the stable equilibrium, and strain A goes extinct. For $1 \lesssim a \lesssim 10$, a stable equilibrium allows for coexistence of both strains, with an increasing equilibrium fraction of strain A. For $10 \lesssim a \lesssim 100$, the previous equilibrium is no longer stable, and undergoes a period doubling bifurcation to produce a stable equilibrium of period 2. As a increases beyond this point, the long term behavior of the system becomes increasely complex, although there are occasional windows of relatively short stable periods. In spite of the chaotic dynamics in this case, we can still see from fig. 5.5 that the two strains will coexist indefinitely as long as the crossing condition is met, with the population fraction in this case varying within the range $x \approx [0.3, 0.8]$.

5.5 Discussion

Varying environmental conditions or frequency-dependent selection appear separately in many models of biological populations. Within the field of population dynamics, changing environments have often been associated with promoting the stable coexistence of strains (16, 2, 30, 8, 5), although chaotic dynamics are also possible (13, 26). Frequency-dependent selection is similarly common in biological models, especially where game theoretic strategies are possible (20, 24), and can also give rise to chaotic dynamics (1, 12).

We present a biologically motivated model of frequency-dependent selection in a periodic environment. Coexistence of both strains is predicted for a wide range of possible fitness assumptions (case II in fig. 5.1). The condition for coexistence is that each strain's relative fitness function is superior to the other's at *some* frequency level, and extinction is only possible when one fitness function dominates the other at *all* frequencies. An experimental consequence of this result is that fitness measurements taken separately in each environmental state for each strain when rare suffice to predict the coexistence or extinction that will result when the strains compete in the periodic environment.

Our model is applicable to viral populations, especially those where defective or complementing viral particles are present. Periodic environmental change is present in commonly studied viral systems, often in the form of alternation between low and high levels of coinfection (31, 37). Models of viral defective particles incorporate frequency dependence (31, 10) and show chaotic behavior similar to what we observe (15).

Our results suggests that stable equilibrium, rather than oscillations or chaotic dynamics, is likely for most biological fitness models of the type studied. Only when the relative fitness of the strains changes very rapidly with frequency does the system exhibit periodic behavior or chaos. We offer a simple mathematical test to exclude the possibility of oscillations or chaos in the system for a given model of frequency dependence.

In arboviruses or other viruses that alternate host species, our results suggest that coexistence between the wild type virus and an undetected mutant strain may be quite common. RNA viruses

are known for their high mutation rate (7), and consequently many different mutants are produced during viral replication. Consider the evolutionarily favored situation of a mutation that confers an advantage in the current host species, but which may have a negative fitness effect in the alternate host. Our results indicate that such a mutant can persist in the wild-type population as long as the negative fitness effects are frequency dependent and not too severe. As such, what may begin as a single wild-type strain may, over time, evolve into two coexisting specialist strains where each adapts well to one host and poorly to the other.

In conclusion, we present a very general model of frequency-dependent selection in a periodic environment. In addition to applications in virology, we expect the generality of our approach will allow this model to be useful broadly within the field of population dynamics and evolutionary ecology.

Bibliography

[1] Altenberg, L. (1991). Chaos from linear frequency-dependent selection. *Am. Nat.*, *138*, 51–68.

[2] Bürger, R. (1999). Evolution of genetic variability and the advantage of sex and recombination in changing environments. *Genetics*, *153*, 1055–1069.

[3] Chen, W.-J., Wu, H.-R., & Chiou, S.-S. (2003). E/NS1 modifications of dengue 2 virus after serial passages in mammalian and/or mosquito cells. *Intervirology*, *46*, 289–295.

[4] Cooper, L. A., & Scott, T. W. (2001). Differential evolution of eastern equine encephalitis virus populations in response to host cell type. *Genetics*, *157*, 1403–1412.

[5] Dean, A. M. (2005). Protecting haploid polymorphisms in temporally variable environments. *Genetics*, *169*, 1147–1156.

[6] Devaney, R. L. (1992). *A First Course in Chaotic Dynamical Systems*. Boston: Addison-Wesley.

[7] Drake, J. W., & Holland, J. J. (1999). Mutation rates among RNA viruses. *Proc. Natl. Acad. Sci. USA*, *96*, 13910–13913.

[8] Edwards, R. J., & Brookfield, J. F. Y. (2003). Transiently beneficial insertions could maintain mobile DNA sequences in variable environments. *Mol. Biol. Evol.*, *20*, 30–37.

[9] Elena, S. F., Miralles, R., & Moya, A. (1997). Frequency-dependent selection in a mammalian RNA virus. *Evolution*, *51*, 984–987.

[10] Frank, S. A. (2000). Within-host spatial dynamics of viruses and defective interfering particles. *J. Theor. Biol.*, *206*, 279–290.

[11] García-Arriaza, J., Manrubia, S. C., Toja, M., Domingo, E., & Escarmís, C. (2004). Evolutionary transition toward defective RNAs that are infectious by complementation. *J. Virol.*, *78*, 11678–11685.

[12] Gavrilets, S., & Hastings, A. (1995). Intermittency and transient chaos from simple frequency-dependent selection. *Proc. R. Soc. Lond. B*, *261*, 233–238.

[13] Hastings, A., Hom, C. L., Ellner, S., Turchin, P., & Godfray, H. C. J. (1993). Chaos in ecology. *Annu. Rev. Ecol. Syst.*, *24*, 1–33.

[14] Huang, A. S. (1973). Defective interfering viruses. *Annu. Rev. Microbiol.*, *27*, 101–117.

[15] Kirkwood, T. B. L., & Bangham, C. R. M. (1994). Cycles, chaos, and evolution in virus cultures: A model of defective interfering particles. *Proc. Natl. Acad. Sci. USA*, *91*, 8685–8689.

[16] Kirzhner, V. M., Korol, A. B., & Ronin, Y. I. (1995). Cyclical environmental changes as a factor in maintaining genetic polymorphism. *J. Evol. Biol.*, *8*, 93–120.

[17] Lenski, R. E., & Velicer, G. J. (2000). Games microbes play. *Selection*, *1*, 89–95.

[18] Li, T.-Y., & Yorke, J. A. (1975). Period three implies chaos. *Am. Math. Monthly*, *82*, 985–992.

[19] Mills, D. R., Peterson, R. L., & Spiegelman, S. (1967). An extracellular darwinian experiment with a self-duplicating nucleic acid molecule. *Proc. Natl. Acad. Sci. USA*, *58*, 217–224.

[20] Mueller, L. D. (1997). Theoretical and empirical examination of density-dependent selection. *Annu. Rev. Ecol. Syst.*, *28*, 269–288.

[21] Novella, I. S., Clarke, D. K., Quer, J., Duarte, E. A., Lee, C. H., Weaver, S. C., Elena, S. F., Moya, A., Domingo, E., & Holland, J. J. (1995). Extreme fitness differences in mammalian and insect hosts after continuous replication of vesicular stomatitis virus in sandfly cells. *J. Virol.*, *69*, 6805–6809.

[22] Novella, I. S., Hershey, C. L., Escarmis, C., Domingo, E., & Holland, J. (1999). Lack of

evolutionary stasis during alternating replication of an arbovirus in insect and mammalian cells. *J. Mol. Biol.*, *287*, 459–465.

[23] Novella, I. S., Reissig, D. D., & Wilke, C. O. (2004). Density-dependent selection in vesicular stomatitis virus. *J. Virol.*, *78*, 5799–5804.

[24] Nowak, M. A., & Sigmund, K. (2004). Evolutionary dynamics of biological games. *Science*, *303*, 793–798.

[25] Perrault, J. (1981). Origin and replication of defective interfering particles. *Curr. Top. Microbiol. Immunol.*, *93*, 151–207.

[26] Rinadli, S., Muratori, S., & Kuznetsov, Y. (1993). Multiple attractors, catastrophes and chaos in seasonally perturbed predator-prey communities. *Bull. Math. Biol.*, *55*, 15–35.

[27] Roux, L., Simon, A. E., & Holland, J. J. (1991). Effects of defective interfering viruses on virus replication and pathogenesis *in vitro* and *in vivo*. *Adv. Virus Res.*, *40*, 181–211.

[28] Sabo, D. L., Domingo, E., Bandle, E. F., Flavell, R. A., & Weissmann, C. (1977). A guanosine to adenosine transition in the 3' terminal extracistronic region of bacteriophage Qb RNA leading to loss of infectivity. *J. Mol. Biol.*, *112*, 235–252.

[29] Sarkovskii, A. N. (1964). Coexistence of cycles of a continuous map of a line into itself. *Ukrain. Mat. Z.*, *16*, 61–71.

[30] Suiter, A. M., Banziger, O., & Dean, A. M. (2003). Fitness consequences of a regulatory polymorphism in a seasonal environment. *Evolution*, *100*, 12782–12786.

[31] Szathmary, E. (1992). Natural selection and dynamical coexistence of defective and complementing virus segments. *J. Theor. Bio.*, *157*, 383–406.

[32] Turner, P. E., & Chao, L. (1998). Sex and the evolution of intrahost competition in RNA virus ϕ6. *Genetics*, *150*, 523–532.

[33] Turner, P. E., & Chao, L. (1999). Prisoner's dilemma in an RNA virus. *Nature*, *398*, 441–443.

[34] Turner, P. E., & Chao, L. (2003). Escape from prisoner's dilemma in RNA phage ϕ6. *Am. Nat.*, *161*, 497–505.

[35] Turner, P. E., & Elena, S. F. (2000). Cost of host radiation in an RNA virus. *Genetics*, *156*, 1465–1470.

[36] Weaver, S. C., Brault, A. C., Kang, W., & Holland, J. J. (1999). Genetic and fitness changes accompanying adaptation of an arbovirus to vertebrate and invertebrate cells. *J. Virol.*, *73*, 4316–4326.

[37] Wilke, C. O., Ressig, D. D., & Novella, I. S. (2004). Replication at periodically changing multiplicity of infection promotes stable coexistence of competing viral populations. *Evolution*, *58*, 900–905.

[38] Yuste, E., Moya, A., & López-Galíndez, C. (2002). Frequency-dependent selection in human immunodeficiency virus type 1. *J. Gen. Virol.*, *83*, 103–106.

[39] Zárate, S., & Novella, I. S. (2004). Vesicular stomatitis virus evolution during alternation between persistent infection in insect cells and acute infection in mammalian cells is dominated by the persistence phase. *J. Virol.*, *78*, 12236–12242.

Appendices

The following appendices contain supplementary information pertaining to Robert Forster's thesis. Appendix A contains the mathematical background and derivations of the statistical tests used in Chapter 2. Appendix B contains additional figures and discussion from the simulations described in Chapter 4. Appendix C provides a mathematical background for the chaotic dynamics described in Chapter 5. Appendix D gives a description of the electronic files provided to the Caltech electronic thesis repository.

Appendix A

Statistical Background for Chapter 2

A.1 The Distribution of the Population's Average Fitness as a Random Variable

In equilibrium, the distribution of the population's average fitness follows from Wright-Fisher sampling. Define π as the probability that a sequence's offspring will be viable. Without resorting to an explicit form for π, equilibrium and a uniform mutation rate imply that all sequences reproduce successfully with the same probability π (which is in general a function of the mutation rate and the mean neutrality of the population). Denote further the expected value of a random variable x as $E[x]$ and its variance as $V[x]$. If we take fitness w_i of the ith offspring in our population as a random variable, the neutral fitness landscape implies that w_i takes only values 0 or 1, where $w_i = 1$ occurs with probability π. The distribution of w_i is therefore a Bernoulli distribution with probability of success π, and we have $E[w_i] = \pi$ and $V[w_i] = \pi(1 - \pi)$.

We now consider the average fitness of the population in equilibrium, $\langle w \rangle$, defined as $\langle w \rangle = \frac{1}{N} \sum_{i=1}^{N} w_i$. By the Central Limit Theorem, the distribution of $\langle w \rangle$ will approach a normal distribution $N[\mu, \sigma^2]$ as $N \to \infty$, and this limit will be reached well before $N = 1000$ (typically $N\pi, N(1 - \pi) > 5$ is sufficient (1), and this condition is easily satisfied under all conditions studied). Thus, $\langle w \rangle$ follows a normal distribution with mean $\mu = E[w] = \pi$ and variance $\sigma^2 = V[w] =$

$\pi(1-\pi)/N$.

To confirm these assumptions hold, we computed the fitness autocorrelation function within a period of equilibrium. Figure A.1 shows the autocorrelation function for the first equilibrium period shown in figure 2.2 (t = 200–9814). The autocorrelation drops almost immediately to a mean of nearly zero, and has a noise level $\sigma \approx 1$–2%, consistent with the variation of w over the time period in question. Similar results hold for each period of fitness equilibrium shown in figure 2.2. In contrast, the population's average neutrality showed significant autocorrelations. While we included

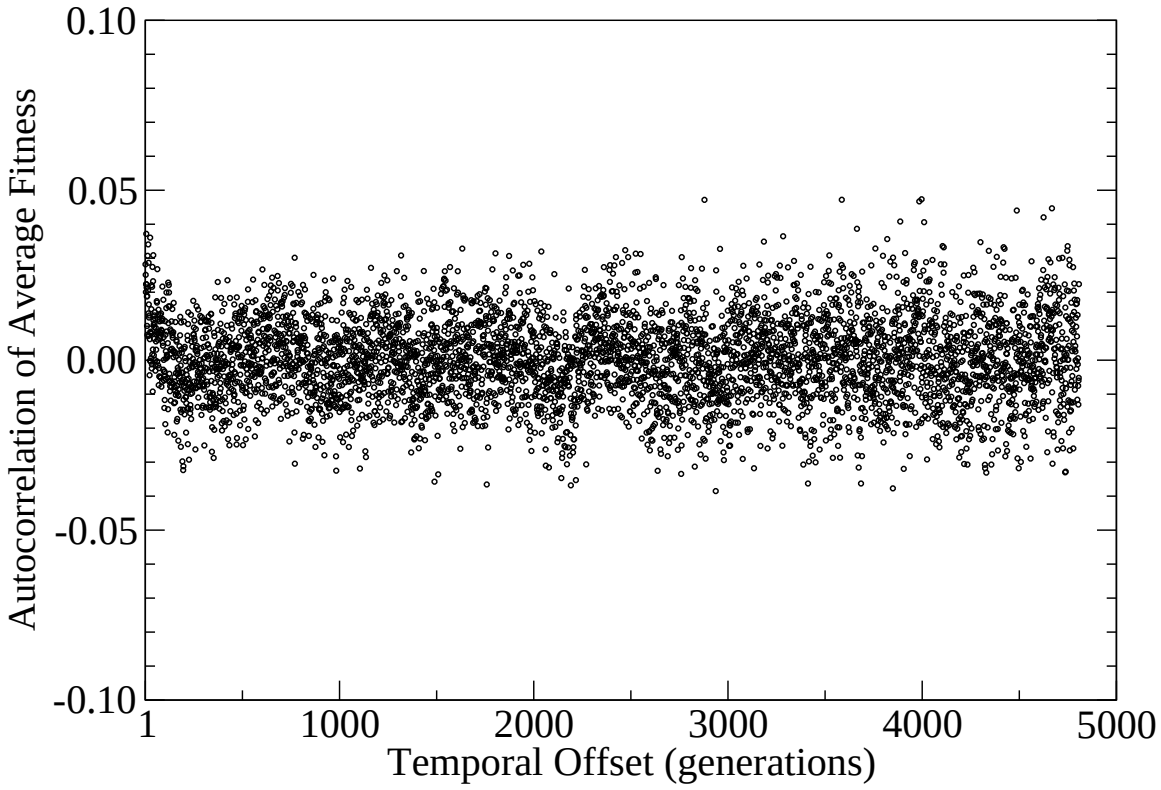


Figure A.1: Temporal autocorrelation function for the first equilibrium period shown in figure 2.2 (t = 200–9814).

A.2 Identifying Jumps in Average Fitness

Motivated by our observations, we seek to characterize the rapid transitions of the population from lower to higher neutrality states. We derive a statistical test for identifying such transitions *a priori* in time series data, and associating a p-value to measure the confidence level of such a transition occurring by chance.

Consider a time series $w(t) \sim N[\mu, \sigma^2]$ measured at T sequential points in time. To test the hypothesis of equal means between two specified time periods $[1, n]$ and $[n+1, T]$ is straightforward, and we will assume equal variances for simplicity. We consider the average value of w over the two periods separately, and consider the sample means Y_i over the two different time periods, defined by $Y_1 = \frac{1}{n}\sum_{t=1}^{n} w(t)$ and $Y_2 = \frac{1}{T-n}\sum_{t=n+1}^{T} w(t)$. These sample means will be normally distributed, $Y_1 \sim N[\mu, \frac{\sigma^2}{n}]$ and $Y_2 \sim N[\mu, \frac{\sigma^2}{T-n}]$. Our null hypothesis is that the means will be equal between the two periods. To test this null hypothesis, we consider the difference between the sample means $D = Y_2 - Y_1$, and ask whether the observed difference can be explained merely by chance, that is, whether the distribution of D is consistent with $D \sim N[0, \sigma_D^2]$. Here, σ_D^2 is the sum of the variances of Y_1 and Y_2, that is, $\sigma_D^2 = \sigma^2 T/[n(T-n)]$.

Thus, under the null hypothesis, the difference of observed means D is normal with zero mean and known variance, and the associated p-value can be obtained by looking up the probability of $Z = D/\sigma_D$ exceeding its observed value in a cumulative distribution table.

We now consider the case of finding the most significant breakpoint in the time series $[1, T]$ when the division into two periods is unspecified. Letting n parameterize the number of data points in the first interval, we can consider the above analysis as a function of n. The highest significance is attained by choosing the maximum value of $D(n)/\sigma_D(n)$, where the difference of means and its variance must be calculated for all n in $[1, T-1]$. Let p_n represent the p-value associated with this maximum n. We wish to know the probability that this maximum level of significance will occur merely by chance due to the fluctuations in $w(t)$. Given $T-1$ independent trials with probability p_n of exceeding our maximum level of significance, we see that the probability of all of these trials

resulting in a smaller significance than that of p_n is

$$\begin{aligned}\Pr[\text{all } T-1 \text{ of the } p_i \text{ satisfy } p_i < p_n] &= (1-p_n)^{T-1} \\ &\approx 1-(T-1)p_n \quad \text{for } p_n \ll 1. \end{aligned} \tag{A.1}$$

From this probability, we calculate the p-value associated with any p_i exceeding our p_n by chance alone, using the above probability:

$$\begin{aligned} p &= \Pr[\text{at least one } p_i \text{ has } p_i > p_n] \\ &= 1 - \Pr[\text{all } T-1 \text{ of the } p_i \text{ satisfy } p_i < p_n] \\ &= 1-(1-p_n)^{T-1} \approx (T-1)p_n\,. \end{aligned} \tag{A.2}$$

Note that the $T-1$ other choices of breakpoints are by no means independent of each other, as they all refer to the same underlying fitness data, $w(t)$. These correlations reduce the number of effective degrees of freedom, and hence the $T-1$ factor will be a conservative overestimate of the actual p-value. If multiple transitions are expected, this algorithm can be repeated on each subinterval to determine whether further breakpoints are consistent with the given level of statistical confidence.

Bibliography

[1] Rice, J. A. (1994). *Mathematical Statistics and Data Analysis*. Duxbury Press, 2nd edn.

Appendix B

Supplemental Materials for Chapter 4

B.1 Additional Simulation Results

The comparison between simulation results and model predictions described in section 4.4.4 was limited by publication constraints. The fused strain varied from $L_{\text{fuse}} = 3$ to 11, although only values $L_{\text{fuse}} = 4, 6, 11$ were shown previously. Figure B.1 and figure B.2 show the remaining results.

B.2 Additional Discussion

Most of the results shown in figures B.1 and B.2 are fairly similar to those shown previously. The limiting cases of $L_{\text{fuse}} = L_{\text{div}} = 5$ and $L_{\text{fuse}} = 2L_{\text{div}} = 10$ are of special interest. These correspond to when the single fused gene can perform both functions with the same length as a single divided gene or with the same length as both divided genes, respectively.

In the first case, $L_{\text{fuse}} = L_{\text{div}} = 5$, the fused strain appears clearly superior. However, looking at figure B.1B, we observe that the outcome of the competition can be random for several combinations of parameters. On the left side of the plot, mutation rates are so low that the extra mutational load of the longer divided strain does not correspond to a significant difference in fitness compared to the fluctuations in the large but finite population. Along the front edge of the plot, the period length has become so long that almost any competition ends due to neutral drift prior to the environment

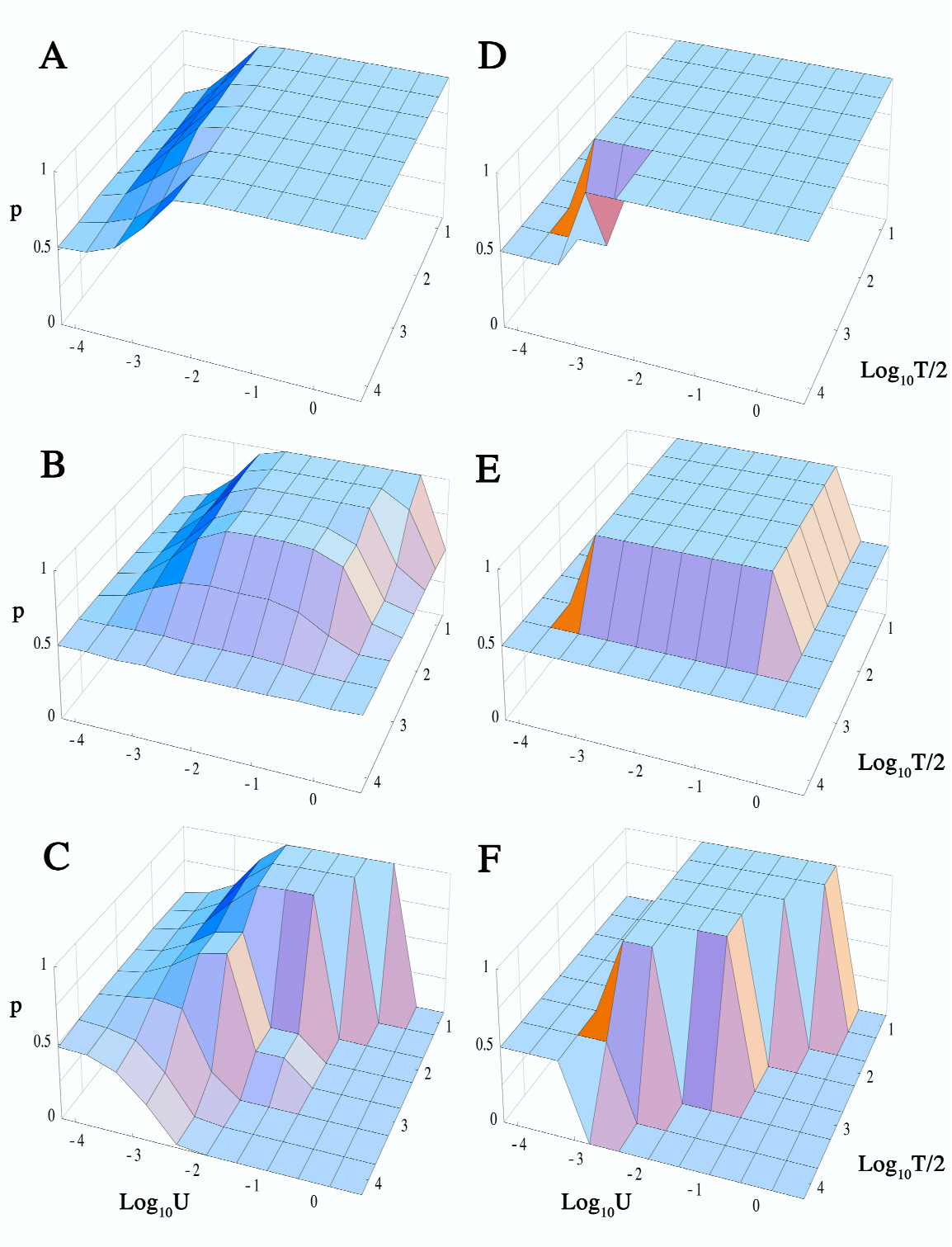


Figure B.1: Left (A, B, C): Simulation results for the probability of fixation of the fused strain as a function of the mutation rate $U_{\text{div}} = \mu L_{\text{div}}$ and period length T for $L_{\text{div}} = 5$ and $L_{\text{fuse}} = 3, 5, 7$ (top to bottom). Simulation results have a standard error of approximately $\pm 1\%$. Right (D, E, F): Model predictions for the same. Parameter values are $s = 1$, $N = 1000$.

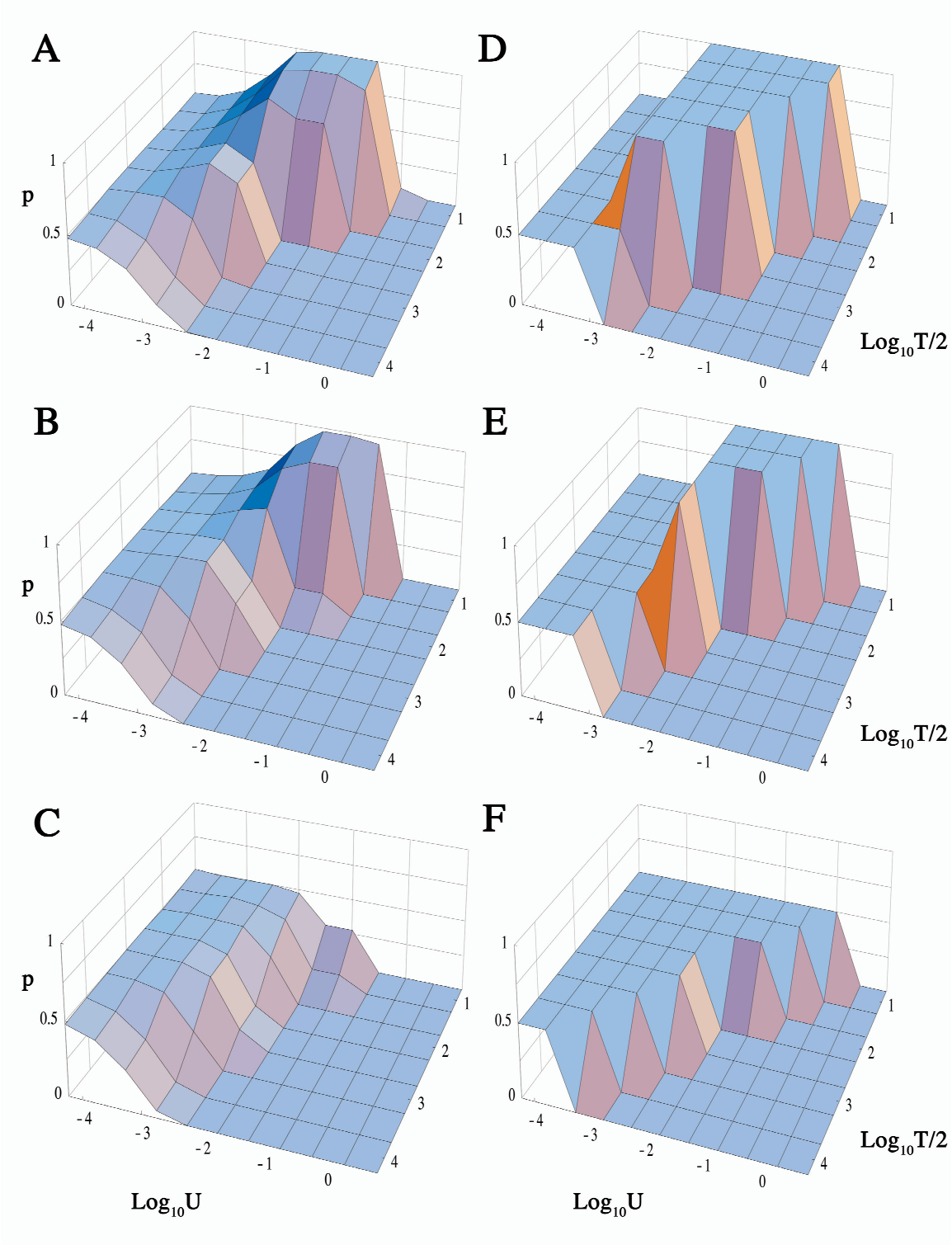


Figure B.2: Left (A, B, C): Simulation results for the probability of fixation of the fused strain as a function of the mutation rate $U_{\text{div}} = \mu L_{\text{div}}$ and period length T for $L_{\text{div}} = 5$ and $L_{\text{fuse}} = 8, 9, 10$ (top to bottom). Simulation results have a standard error of approximately $\pm 1\%$. Right (D, E, F): Model predictions for the same. Parameter values are $s = 1$, $N = 1000$.

changing even once ($T_{\mathrm{dr}} \ll T$). Clearly in this case the two strains are equivalent and the second gene in the divided strain plays no role. Lastly, on the right edge of the plot the mutation rate is so high that the population is effectively randomized over all possible strings. The combinatorial fraction of each strain that is functional is the same since the environmentally favored gene has length 5 for both strains.

In the second case, $L_{\mathrm{fuse}} = 2L_{\mathrm{div}} = 10$, the fused strain appears clearly inferior. Looking at figure B.2C, the outcome is only random at very low mutation rates where the additional length of the fused genes relative to those of the divided strain are unimportant. However, note that the fused strain fares somewhat less worse at high mutation rates and short periods. This difference is due to a dynamic effect that was neglected in our model for simplicity. Recall that the fused strain quickly reaches its equilibrium quasispecies distribution (see figure 4.1) and remains there independent of the environmental changes. In contrast, the distribution of the divided strain must re-adjust every time the environment changes. At short to intermediate period lengths, the divided strain will begin to lose its previously functional but currently useless gene to neutral drift, and will lose it more quickly at high mutation rates. If the period is reasonably short, the divided strain is constantly re-adjusting to the new environment and does not have as much time to enjoy its equilibrium advantage over the fused strain before the environment changes again. This is the origin of the dynamic effect and its importance depends on the *quasispecies timescale*, T_{q}, the time that the divided strain takes to mostly reach its new equilibrium distribution after an environmental change.

Appendix C

Supplemental Materials for Chapter 5

C.1 Mathematical Background: Finding All Attracting Orbits

To understand the dynamics of our system, several results from the mathematics of one-dimensional iterated maps are relevant (1, 3). A useful tool in understanding the long-term behavior of the system is the Schwarzian derivative of a function $g(x)$, $Sg(x)$:

$$Sg(x) = \frac{g'''(x)}{g'(x)} - \frac{3}{2}\left(\frac{g''(x)}{g'(x)}\right)^2 . \tag{C.1}$$

In the study of iterated maps, functions with strictly negative Schwarzian derivatives, i.e., $\forall x\, Sg(x) < 0$, are mathematically well behaved. In particular, when this condition holds there exists a critical point of $g(x)$ [where $dg(x)/dx = 0$] whose orbit is eventually attracted to any attracting periodic point of the system (2). Therefore if the Schwarzian derivative is strictly negative, only the critical points of $g(x)$ must be studied to determine the nature of the attracting orbits of the system.

In our chaotic example using the fitness functions given in eq. (5.11), once the parameter a exceeds $a_c \approx 10^{0.87}$, we have $Sg(x) < 0$ for all x. This critical value of a_c can be found either by use of the Schwarzian derivative condition or by solving the stability condition given in eq. (5.10).

For a larger than the critical value a_c, $g(x)$ has two critical points. These critical points are found numerically and both of their orbits are plotted after a large number of iterations in fig. 5.5. Since these critical points converge to an attracting fixed point of the system, we can see all the possible long-term behaviors of the system by examining only the orbits of these points.

For values of a less than the cutoff value a_c, there are no critical points of $g(x)$ as $dg(x)/dx$ is strictly positive. This implies that $g(x)$ is an increasing function, and, as in the linear case, this rules out any periodic behavior. For these smaller values of $a < a_c$, we arbitrarily plot the orbit of $x = 1/2$ in fig. 5.5.

C.2 The Effect of Noise on Equilibrium

We consider the case that the fitness functions $w_A(x)$ and $w_B(x)$ given in eq. (5.11) may be "noisy" rather than exactly known. We model the noise component of these functions as a multiplicative factor of $e^{\beta z}$ where z is a gaussian random variable with mean zero and variance 1. The parameter β sets the scale of the noise and is typically small compared to unity. Figures C.1 and C.2 show the long-term population dynamics in the presence of noise levels $\beta = 0.001$ and 0.01 respectively. These levels correspond to a typical variation in fitness of 0.1% and 1% respectively. All other conditions are identical to those used in fig. 5.5.

The addition of a small noise component obscures the longer periodic orbits in the system. While in the noiseless system chaotic behavior requires a fairly rapid variation in w_B as a function of x, a small amount of noise causes the system to descend into apparent chaos. With only the moderately steep changes in $w_B(x)$ given by $a = 10^2$ (see fig. 5.2), the system becomes essentially unpredictably when there are 1% variations in fitness as shown in figure C.2. Smaller levels of noise such as those shown in figure C.1 preserve most of the periodic structure of the equilibria although very high period orbits become effectively chaotic under these conditions.

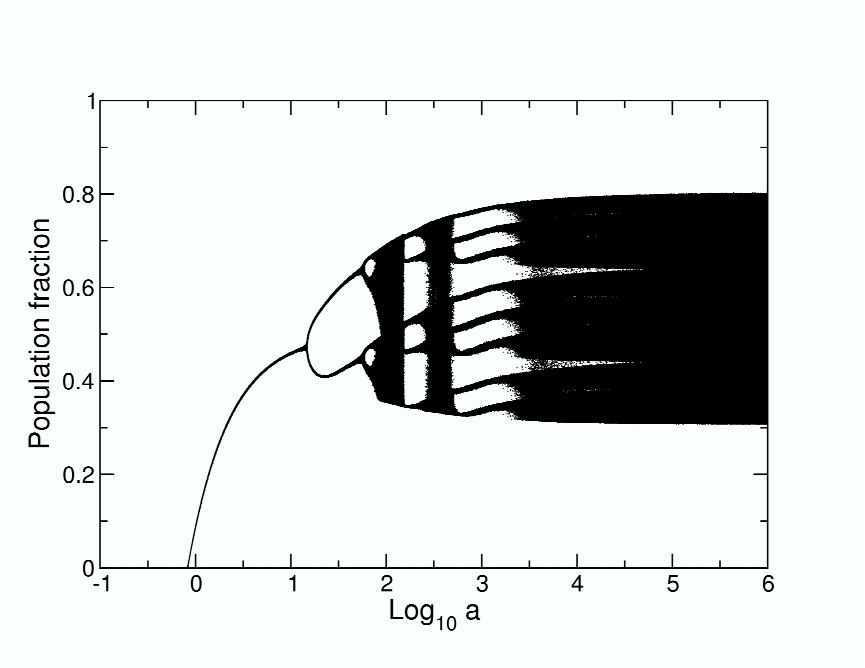


Figure C.1: Population fraction in the long term, plotted as a function of the parameter a in eq. (5.11). The noise level in the underlying fitness functions is $\beta = 0.001$, corresponding to a typical variation in the fitness of 0.1%. All other assumptions are the same as in fig. 5.5.

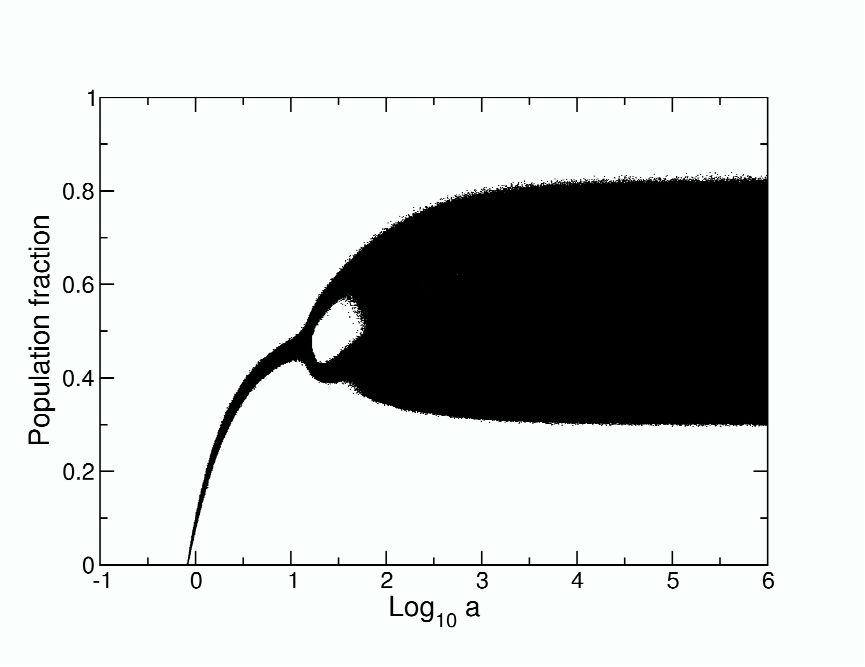


Figure C.2: Population fraction in the long term, plotted as a function of the parameter a in eq. (5.11). The noise level in the underlying fitness functions is $\beta = 0.01$, corresponding to a typical variation in the fitness of 1%. All other assumptions are the same as in fig. 5.5.

Bibliography

[1] Beardon, A. F. (1991). *Iteration of Rational Functions*. New York: Springer-Verlag.

[2] Devaney, R. L. (1992). *A First Course in Chaotic Dynamical Systems*. Boston: Addison-Wesley.

[3] Georgescu, C., Joita, C., Nowell, W. O., & Stanica, P. (2005). Chaotic dynamics of some rational maps. *Discrete and Continuous Dyn. Systems*, *12*, 363–375.

Appendix D

Description of Electronic Files

The electronic version of this thesis contains a number of additional files as described below. You may access these electronic files from the Caltech electronic thesis repository. Please see http://etd.caltech.edu/ for more information.

- chapter2.tgz - a zipped tar archive of the programs and source code used to generate the results in chapter 2. This includes the C++ programs and Perl scripts used to perform and process the output of the simulation illustrated in figure 2.2, perform the statistical step-finding and the autocorrelation analysis described in appendix A, and perform the sampling of RNA neutral networks needed to estimate their size distribution. Also included are Microsoft Excel files analyzing and summarizing the step statistics and consensus sequence evolution. The data for figure 2.2 are also included since the neutrality data shown took an entire supercomputer cluster several weeks to generate, a situation hopefully soon to appear dated. The RNA folding algorithms used were those of the Vienna RNA package, version 1.4, which is also included.

- chapter4.tgz - a zipped tar archive of the programs and source code used to generate the results in both chapters 3 and 4. In particular, this includes C++ programs that

  - calculate the quasispecies distributions of either the fused or divided strains
  - perform fixation experiments to see how often and how long it takes for a single member of one strain to successfully invade a population of the other strain
  - perform injection experiments to see how long it takes for one strain to successfully invade

a population other when the invading strain has a single individual injected initially and again as soon as the invading population goes extinct

- perform competition simulations between the fused and divided strains across a range of parameter values and determine the probability that the fused strain goes to fixation

Also included are a Microsoft Excel file that implement the model described in section 4.4.3 and the Mathematica notebook used to generate figures 4.2, B.1, and B.2.

- chapter5.tgz - a zipped tar archive of the programs and source code used to generate the results in chapter 5. In particular, this includes the C++ program used to generate the data shown in the plot of the chaotic dynamics shown in figure 5.5 and a Mathematica notebook used to produce the data for the other figures in chapter 5.