EVOLUTIONARY DYNAMICS

EXPLORING THE EQUATIONS OF LIFE

MARTIN A. NOWAK

Printed in Canada

Library of Congress Cataloging-in-Publication Data

Nowak, M. A. (Martin A.)
Evolutionary dynamics : exploring the equations of life /
Martin A. Nowak.
p. cm.
Includes bibliographical references and index.
ISBN-13: 978-0-674-02338-3 (alk. paper)
ISBN-10: 0-674-02338-2 (alk. paper)
1. Evolution (Biology)—Mathematical models. I. Title.

QH371.3.M37N69 2006
576.801'5118—dc22 2006042693

Designed by Gwen Nefsky Frankfeldt

PREFACE

Evolutionary Dynamics presents those mathematical principles according to which life has evolved and continues to evolve. Since the 1950s biology, and with it the study of evolution, has grown enormously, driven by the quest to understand the world we live in and the stuff we are made of. Evolution is the one theory that transcends all of biology. Any observation of a living system must ultimately be interpreted in the context of its evolution. Because of the tremendous advances over the last half century, evolution has become a discipline that is based on precise mathematical foundations. All ideas regarding evolutionary processes or mechanisms can, and should, be studied in the context of the mathematical equations of evolutionary dynamics.

The original formulation of evolutionary theory and many of the investigations of its first hundred years dealt with the genetic evolution of the origin and adaptation of species. But more recently evolutionary thinking has expanded to all areas of biology and many related disciplines of the life sciences. Wherever information reproduces, there is evolution. Mutations are caused by errors in information transfer, resulting in different types of messages. Selection among types emerges when some messages reproduce faster than others.

Mutation and selection make evolution. Mutation and selection can be described by exact mathematical equations. Therefore evolution has become a mathematical theory.

The life sciences in general, and biology in particular, are on the brink of an unprecedented theoretical expansion. Every university is currently aiming to establish programs in mathematical biology and to offer its students an interdisciplinary education that spans fields as diverse as mathematics and molecular biology, linguistics and computer science. At the borders of such disciplines, progress occurs. Whenever the languages of two disciplines meet, two cultures interact, and something new happens.

In this book, the languages of biology and mathematics meet to talk about evolution. *Evolutionary Dynamics* introduces the reader to the fascinating and simple laws that govern the evolution of living systems, however complicated they may seem. I will start with the basics, avoid unnecessary complications, and reach cutting-edge research problems within a few steps.

The book grew out of a course I taught at Harvard University in 2004 and 2005. The students in my first class were Blythe Adler, Natalie Arkus, Michael Baym, Paul Berman, Illya Bomash, Nathan Burke, Chris Clearfield, Rebecca Dell, Samuel Ganzfried, Michael Gensheimer, Julia Hanover, David Hewitt, Mark Kaganovich, Gregory Lang, Jonathan Leong, Danielle Li, Alex Macalalad, Shien Ong, Ankit Patel, Yannis Paulus, Jura Pintar, Esteban Real, Daniel Rosenbloom, Sabrina Spencer, and Martin Willensdorfer, and the teaching fellows Erez Lieberman, Franziska Michor, and Christine Taylor. I have learned much from you. Your questions were my motivation. I wrote this book for you.

I am indebted to many people. Most of all I would like to thank May Huang and Laura Abbott, who helped me to prepare the final manuscript and index. They turned chaos into order. I could not have finished without them. I also thank the excellent editors of Harvard University Press, Elizabeth Gilbert and Michael Fisher.

I thank Ursula, Sebastian, and Philipp for their patience and for their burning desire to understand everything that can be understood.

I would like to express my gratitude to my teachers, Karl Sigmund and Robert May. Both of them are shining examples of how scientists should be. They have again and again impressed me with their superior judgment, in-

sight, and generosity. I also appreciate the work and friendship of the many people with whom I have had the honor of collaborating and whose enthusiasm for science is woven into the ideas presented here: Roy Anderson, Rustom Antia, Ramy Arnaout, Charles Bangham, Barbara Bittner, Baruch Blumberg, Maarten Boerlijst, Sebastian Bonhoeffer, Persephone Borrow, Reinhard Bürger, Michael Doebeli, Peter Doherty, Andreas Dress, Ernst Fehr, Steve Frank, Drew Fudenberg, Beatrice Hahn, Christoph Hauert, Tim Hughes, Lorens Imhof, Yoh Iwasa, Vincent Jansen, Paul Klenerman, Aron Klug, Natalia Komarova, David Krakauer, Christoph Lengauer, Richard Lenski, Bruce Levin, Erez Lieberman, Jeffrey Lifson, Marc Lipsitch, Alun Lloyd, Joanna Masel, Erick Matsen, Lord May of Oxford (Defender of Science), John Maynard Smith, Angela McLean, Andrew McMichael, Franziska Michor, Garrett Mitchener, Richard Moxon, Partha Niyogi, Hisashi Ohtsuki, Jorge Pacheco, Karen Page, Robert Payne, Rodney Phillips, Joshua Plotkin, Roland Regoes, Ruy Ribeiro, Akira Sasaki, Charles Sawyers, Peter Schuster, Anirvan Sengupta, Neil Shah, George Shaw, Karl Sigmund, Richard Southwood, Ed Stabler, Dov Stekel, Christine Taylor, David Tilman, Peter Trappa, Arne Traulsen, Bert Vogelstein, Lindi Wahl, Martin Willensdorfer, and Dominik Wodarz.

I thank Jeffrey Epstein for many ideas and for letting me participate in his passionate pursuit of knowledge in all its forms.

INTRODUCTION

IN 1831, at the age of twenty-two, Charles Darwin embarked on his journey around the world. He gazed at the breath-taking diversity of tropical flora and fauna, collected creepy-crawlies from the vast oceans that he traversed, was hopelessly seasick, saw slavery in Brazil, witnessed genocide in Argentina, and was underwhelmed by the naked humanity at Tierra del Fuego. He experienced the effects of a devastating earthquake in Chile that raised the South American continent. He led an expedition into the Andes and discovered marine fossils at high altitude. He paid little attention to which finches came from which islands in the Galápagos and ate most of the delicious turtles he had gathered on his way home across the Pacific. He saw Tahiti and the economic rise of Australia. He visited John Hershel, England's leading physicist of the time, in South Africa; Hershel told him that "the mystery of mysteries" was the as yet unknown mechanism that gave rise to new species. Darwin returned to England's shores after five years, having collected six thousand specimens that would require decades of analysis by an army of experts.

His own observations in geology and the theory of his mentor, Sir Charles Lyell, that mountains were not lifted up in one day, but rose slowly over

unimaginable periods of time, led Darwin to a key idea: given enough time everything can happen.

Charles Darwin did not invent the concept of evolution. When he was a student in Edinburgh in the late 1820s, evolution was already the talk of the town. But evolution was rejected by the establishment. Those who adhered to evolutionary thinking were called Lamarckists, after the French scientist Jean-Baptiste Lamarck, who was the first to propose that species are not static, but change over time and give rise to new species. Lamarck had offered this perspective in a book published in 1809. He did not, however, propose a correct mechanism for how species change into each other. This mechanism was discovered first by Charles Darwin and independently by Alfred Russel Wallace.

From reading the economist Thomas Malthus, Darwin was aware of the consequences of exponentially growing populations. Once resources become limiting only a fraction of individuals can survive. Darwin was also a keen observer of animal breeders. He analyzed their methods and studied their results. Slowly he understood that nature acted like a gigantic breeder. This was the first time that natural selection materialized as an idea, a scientific concept in a human mind. Darwin was thirty-three years old.

The one problem that Darwin did not solve concerned the mechanism that could maintain enough diversity in a population for natural selection to operate. Darwin was unaware of the Austrian monk and botanist Gregor Mendel and his experiments on plant heredity. Mendel's work had already been published but was hidden, gathering dust in the *Annals* of the Brno Academy of Sciences.

Darwin once remarked, "I have deeply regretted that I did not proceed far enough at least to understand something of the great leading principles of mathematics; for men thus endowed seem to have an extra sense." The engineer Fleeming Jenkins, who reviewed Darwin's *On the Origin of Species*, published in 1859, had raised a fundamental and seemingly intractable objection to Darwin's theory: if offspring inherit a blend of the parents' characteristics, then variability diminishes in successive generations. Several decades later a simple mathematical equation, independently found by the famous British mathematician G. H. Hardy and the German physician Wilhelm Weinberg, showed that Mendelian (particulate) inheritance does lead to a maintainance

of genetic diversity under random mating. The Hardy-Weinberg law is one of the fundamental principles of evolution under sexual reproduction.

Mendelian genetics and Darwinian evolution were unified in the new discipline of mathematical biology, which developed from the seminal investigations of Ronald Fisher, J. B. S. Haldane, and Sewall Wright in the 1920s and 1930s. Through their work, fundamental concepts of evolution, selection, and mutation were embedded in a precise mathematical framework. This line of mathematical analysis was taken up in the 1950s by Motoo Kimura, who formulated the neutral theory of evolution. Kimura realized that most genetic mutations do not affect fitness and are fixed in populations only by random drift.

Other milestones of evolutionary dynamics include William Hamilton's discovery in 1964 that selection of "selfish genes" can favor altruistic behavior among relatives and John Maynard Smith's invention of evolutionary game theory in 1973. In the mid-1970s Robert May revolutionized the mathematical approaches to ecology and epidemiology. Manfred Eigen and Peter Schuster formulated quasispecies theory, which provides a link between genetic evolution, physical chemistry, and information theory. Peter Taylor, Josef Hofbauer, and Karl Sigmund studied the replicator equation, the foundation of evolutionary game dynamics.

This very brief and incomplete account of the evolution of evolutionary dynamics brings us to the present book. It has fourteen chapters. Although there is some progression of complexity, the chapters are largely independent. Therefore, if you know something about the subject, you can read the book in whatever order you like. My aim has been to keep things as simple as possible, as linear as possible, and as deterministic as possible. I will start with the basics and in a few steps lead you to some of the most interesting and unanswered research questions in the field. Having read the book, you will know what you need to embark on your own journey and make your own discoveries.

This book represents an introduction to certain aspects of mathematical biology, but it is not comprehensive. Mathematical biology includes many topics, such as theoretical ecology, population genetics, epidemiology, theoretical immunology, protein folding, genetic regulatory networks, neural networks, genomic analysis, and pattern formation. The field is too diverse for any one book to represent it without running the risk of becoming as entertaining as a

telephone directory. I have chosen those topics that I know well and where my explanation can be brief and effective. I have concentrated on evolution because it is the one unifying principle of all of biology.

It might seem surprising that a book on evolutionary dynamics is not primarily about population genetics. Nevertheless the ideas and concepts of this fascinating field stand behind many of my explorations: the basic mathematical formulations of selection, mutation, random drift, fitness landscapes, and frequency-dependent selection as well as of evolution in structured populations have originated in population genetics. Several major themes of population genetics, however, such as sexual reproduction, sexual selection, recombination, and speciation, are not discussed here. In contrast, classical population genetics does not deal with evolutionary dynamics of infectious agents, the somatic evolution of cancer, evolutionary game theory, or the evolution of human language, all of which are subjects that I do explore.

The main ingredients of evolutionary dynamics are reproduction, mutation, selection, random drift, and spatial movement. Always keep in mind that the population is the fundamental basis of any evolution. Individuals, genes, or ideas can change over time, but only populations evolve.

The structure of the book is as follows. After this introduction, in Chapter 2 I will discuss populations of reproducing individuals and the basic ideas of natural selection and mutation. Simple models of population dynamics can lead to an exponential explosion, to a stable equilibrium, or to oscillations and chaos. Selection emerges whenever two or more individuals reproduce at different rates. Mutation means that one type can change into another. There are models of population growth that lead to the survival of whoever reproduces fastest ("survival of the fittest"). Other models lead to the survival of the first or the coexistence of all.

In Chapter 3, quasispecies theory is introduced. Quasispecies are populations of reproducing genomes subject to mutation and selection. They live in sequence space and move over fitness landscapes. An important relationship between mutation rates and genome length is called the "error threshold": adaptation on most fitness landscapes is possible only if the mutation rate per base is less than one over the genome length, measured in bases.

In Chapter 4, we study evolutionary game dynamics, which arise whenever the fitness of an individual is not constant but depends on the relative

abundance (= frequency) of others in the population. Thus evolutionary game theory is the most comprehensive way to look at the world. People who do not engage in evolutionary game theory restrict themselves to the rigidity of constant selection, where the fitness of one individual does not depend on others. The replicator equation is a nonlinear differential equation that describes frequency-dependent selection among a fixed number of strategies. We will encounter the Nash equilibrium and evolutionarily stable strategies. Evolutionary game theory and ecology are linked in an important way: the replicator equation is equivalent to the Lotka-Volterra equation of ecological systems, which describes the interation between predator and prey species.

Chapter 5 is dedicated to the best game in town, the Prisoner's Dilemma. The cooperation of reproducing entities is essential for evolutionary progress. Genes cooperate to form a genome. Cells cooperate to produce multicellular organisms. Individuals cooperate to form groups and societies. The emergence of human culture is a cooperative enterprise. The very problem of how to obtain cooperation by natural selection is described by the Prisoner's Dilemma. In the absence of any other assumption, natural selection favors defectors over cooperators. Cooperation has a chance, however, if there are repeated interactions between the same two individuals. We will encounter the strategy Tit-for-tat, which is defeated first by Generous Tit-for-tat and then by Win-stay, lose-shift.

In Chapter 6 we move to a stochastic description of finite populations. Neutral drift is a crucial aspect of evolutionary dynamics: if a finite population consists of two types of individuals, red and blue, and if both individuals have identical fitness, then eventually the population will be either all red or all blue. Even in the absence of selection, coexistence is not possible. If there is a fitness difference, then the fitter type has a greater chance of winning, but no certainty. We calculate the probability that the descendants of one individual will take over the whole population. This so-called fixation probability is important for estimating the rate of evolution.

Chapter 7 is about games in finite populations. Most of evolutionary game theory has been formulated in terms of deterministic dynamics describing the limit of infinitely large populations. Here we move game theory to finite populations and make surprising observations. Neither a Nash equilibrium, nor an evolutionarily stable strategy, nor a risk-dominant strategy is protected

WHAT EVOLUTION IS

THIS CHAPTER introduces three basic building blocks of evolutionary dynamics: replication, selection, and mutation. These are the fundamental and defining principles of biological systems. They apply to any biological organization anywhere in our or other universes and do not depend on the particular details of which chemistry was recruited to embody life. Any living organism has arisen and is continually modified by these three principles.

Evolution requires populations of reproducing individuals. In the right environment, biological entities, such as viruses, cells, and multicellular organisms can make copies of themselves. The blueprint that determines their structure, the genomic material in form of DNA or RNA, is replicated and passed on to the offspring. Selection results when different types of individuals compete with each other. One type may reproduce faster and thereby outcompete the others. Reproduction is not perfect, but involves occasional mistakes, or mutations. Mutation is responsible for generating different types that can be evaluated in the selection process, and thus results in biological novelty and diversity. Selection will choose to maintain some innovations and dismiss others, and can favor or oppose genetic diversity.

At the end of this chapter we will focus on the Hardy-Weinberg law of random mating. This discussion will be our only venture into the mathematics of sexual reproduction. In subsequent chapters we will encounter additional principles of evolutionary dynamics, such as random drift and spatial movement.

2.1 REPRODUCTION

Imagine a single bacterial cell in a perfect environment that contains all the nutrients required for growth and happiness. In this bacterial heaven, the fortunate cell and all its offspring divide every 20 minutes, which is the known world record for bacterial cell division in an ideal lab setting. After 20 minutes the cell has given rise to 2 daughter cells. After 40 minutes there are 4 granddaughters, and after one hour there are 8 great granddaughters. How many cells will there be after three days?

After t generations there are 2^t cells. In three days there are 216 generations. Hence we expect $2^{216} = 10^{65}$ cells. The total mass of these cells would exceed the mass of the earth by many orders of magnitude.

The growth law for this overwhelming expansion can be written as a recursive equation

$$x_{t+1} = 2x_t. \tag{2.1}$$

Here x_t is the number of cells at time t, and x_{t+1} is the number of cells at time $t + 1$. The equation means that at time $t + 1$ there are twice as many cells as at time t. Time is measured in numbers of generations.

The number of cells at time 0 is given by x_0. With this initial condition, the solution of equation (2.1) can be written as

$$x_t = x_0 2^t. \tag{2.2}$$

Equation (2.1) is a so-called difference equation, because time is measured in discrete steps.

We can also formulate a differential equation for exponential growth that measures time as a continuous quantity. Let $x(t)$ denote the abundance of cells at time t. Suppose that cells divide at rate r. More precisely, we assume that the

time for cell division follows an exponential distribution with average $1/r$. We can write the differential equation

$$\dot{x} = \frac{dx}{dt} = rx. \tag{2.3}$$

Throughout this book, I will use the standard notation $\dot{x}$ to refer to differentiation (of x) with respect to time. If the abundance of cells at time 0 is given by x_0 then the solution of the differential equation (2.3) is

$$x(t) = x_0 e^{rt}. \tag{2.4}$$

Let us reconsider our bacterial supernova. If we measure time in units of days, then $r = 72$ means that the time for a cell cycle requires, on average, 20 minutes (calculated by dividing the total number of minutes in a day, 1,440, by 72). Hence there are 72 cell divisions in one day. After three days, one bacterial cell has generated e^{216} cells which is approximately 6×10^{93} cells.

The discrepancy between the differential equation and the difference equation is a consequence of the varying assumptions for the distribution of the generation time. The difference equation assumes that each cell division occurs after exactly 20 minutes. The differential equation assumes that each cell division occurs after a time which is exponentially distributed around an average of 20 minutes. The exponential distribution is defined as follows: the probability that cell division occurs between time 0 and τ is given by $1 - e^{-r\tau}$. On average, cells divide after $1/r$ time units.

So far we have ignored cell death. Let us now suppose that cells die at rate d, which means that they have an exponentially distributed lifespan with an average of $1/d$. The differential equation becomes

$$\dot{x} = (r - d)x. \tag{2.5}$$

The effective growth rate is the difference between the birth rate, r, and the death rate, d. If $r > d$, then the population will expand indefinitely. If $r < d$, then the population will converge to zero and become extinct. If $r = d$, then the population size remains constant, but this situation is unstable: small deviations from absolute equality between birth and death will lead to either exponential expansion or decline. It is important to note that setting $r = d$

in equation (2.5) does not constitute a mechanism for maintaining a stable constant population size.

The simple equation (2.5) allows us to introduce an extremely important concept in evolution, ecology and epidemiology: the basic reproductive ratio, r/d. This ratio denotes the expected number of offspring that come from any one individual. The average lifetime of a cell is $1/d$. The rate of producing offspring cells is given by r. If each cell produces on average more than one offspring, $r/d > 1$, then an exponential expansion will follow. A basic reproductive ratio greater than one is a necessary condition for population expansion.

We have observed that ongoing exponential growth can lead to unreasonably high numbers in a very short time. In a realistic environment, the expanding population will hit constraints that prevent further expansion. For example, the population might run out of nutrients or physical space.

A model for population expansion with a maximum carrying capacity is given by the logistic equation

$$\dot{x} = rx(1 - x/K). \tag{2.6}$$

As before, the parameter r refers to the rate of reproduction in the absence of density regulation, when the population size, x, is much smaller than the carrying capacity K. As x increases, the rate of growth slows down. When x reaches the carrying capacity, K, then the population expansion ceases. For the initial condition x_0, the solution of equation (2.6) is given by

$$x(t) = \frac{Kx_0e^{rt}}{K + x_0(e^{rt} - 1)}. \tag{2.7}$$

In the limit of infinite time, $t \to \infty$, the population size converges to the equilibrium $x^* = K$. Throughout the book we will use a superscript asterisk to denote a quantity at equilibrium.

2.1.1 Deterministic Chaos

We can also study a logistic difference equation. Without loss of generality, let us rescale the population abundance in such a way that the maximum carrying capacity is given by $K = 1$. We have

Hence the A and B subpopulations grow exponentially at rates a and b, respectively. The doubling time for A is $\log 2/a$. The doubling time for B is $\log 2/b$. If a is greater than b, then A reproduces faster than B: after some time, there will be more A than B individuals.

Denote by $\rho(t) = x(t)/y(t)$ the ratio of A over B at time t. We have

$$\dot{\rho} = \frac{\dot{x}\,y - x\,\dot{y}}{y^2} = (a - b)\rho. \tag{2.11}$$

The solution of this differential equation, for the initial condition $\rho_0 = x_0/y_0$, is given by

$$\rho(t) = \rho_0 e^{(a-b)t}. \tag{2.12}$$

Hence if $a > b$ then ρ tends to infinity. In this case A will outcompete B, which means selection favors A over B. If, on the other hand, $a < b$, then ρ tends to zero. In this case B will outcompete A, which means that selection favors B over A.

Let us now consider a situation in which the total population size is held constant. This situation can arise, for example, when an ecosystem has a constant maximum carrying capacity. Let $x(t)$ denote the relative abundance of A at time t. Instead of "relative abundance" we can also say "frequency." Let $y(t)$ denote the frequency of B. Since there are only A and B individuals in the population, we have $x + y = 1$. As before, A and B individuals reproduce, respectively, at rates a and b.

We have the system of equations

$$\begin{aligned} \dot{x} &= x(a - \phi) \\ \dot{y} &= y(b - \phi) \end{aligned} \tag{2.13}$$

The term ϕ ensures that $x + y = 1$. This is only possible if $\phi = ax + by$. Observe that ϕ is the average fitness of the population.

The system (2.13) describes only a single differential equation, because y can be replaced by $1 - x$. We obtain

$$\dot{x} = x(1 - x)(a - b). \tag{2.14}$$

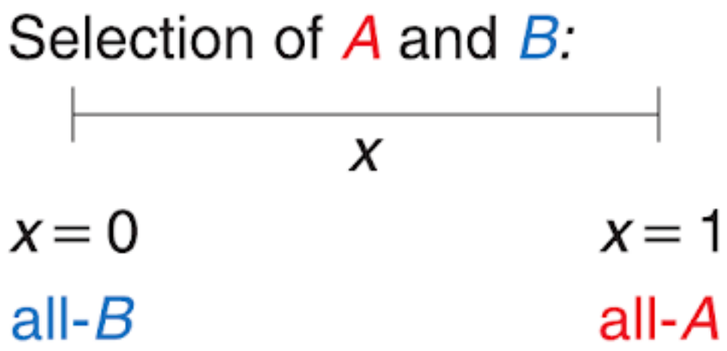


Figure 2.2 Selection arises if two types, A and B, have different rates of reproduction, a and b. If A reproduces faster than B, which means $a > b$, then A will become more abundant than B. Eventually A will take over the entire population; B will become extinct. Denote by x the relative abundance (= frequency) of type A. The quantity x is a number between 0 and 1. Therefore selection dynamics are defined on the closed interval $[0, 1]$.

This differential equation has two equilibria, one for $x = 0$ and the other for $x = 1$. At these two points, we have $\dot{x} = 0$. This observation makes sense: if $x = 1$ then the system consists only of A individuals and nothing more can happen; if $x = 0$, then the system consists only of B individuals and again nothing more can happen.

We can, however, make an additional observation. If $a > b$, then $\dot{x} > 0$ for all values of x that are strictly greater than 0 and strictly smaller than 1. This means that for any mixed system (consisting of some A and some B individuals) the fraction of A will increase if the fitness of A is greater than the fitness of B. In this case, the fraction of B will converge to 0, while the fraction of A converges to 1. We have encountered the concept of "survival of the fitter" (Figure 2.2).

2.2.1 Survival of the Fittest

The model can be extended to describe selection among n different types. Let us label them $i = 1, \ldots, n$. Denote by $x_i(t)$ the frequency of type i. The structure of the population is given by the vector $\vec{x} = (x_1, x_2, \ldots, x_n)$.

Denote by f_i the fitness of type i. As before, fitness is a non-negative real number and describes the rate of reproduction. The average fitness of the

The simplex is the set of all points whose coordinates are not negative and add up to one

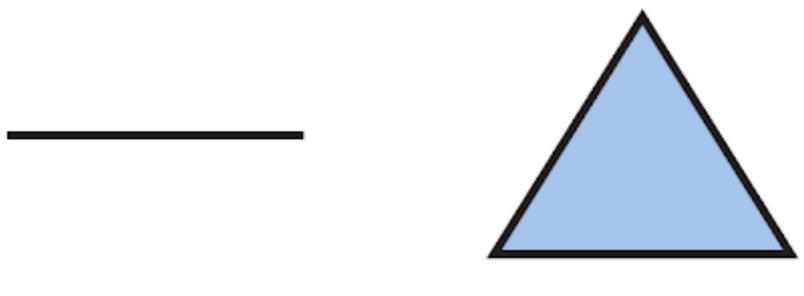


Figure 2.3 If the total population size is constant, then selection dynamics can be formulated in terms of relative abundance (= frequency). Suppose there are n different types, $i = 1, \ldots, n$. Type i has frequency x_i. The sum over all x_i is one. The set of all points, $(x_1, \ldots, x_n)$ with the property $\sum_{i=1}^{n} x_i = 1$, is called the simplex S_n. Selection dynamics occur on the simplex S_n. The figure shows S_2, S_3, and S_4. The simplex S_n is an $n - 1$ dimensional structure embedded in an n-dimensional Euclidian space. The simplex S_n has n faces that each consist of the simplex S_{n-1}.

population is given by

$$\phi = \sum_{i=1}^{n} x_i f_i . \tag{2.15}$$

Selection dynamics can be written as

$$\dot{x}_i = x_i(f_i - \phi) \qquad i = 1, \ldots, n \tag{2.16}$$

The frequency of type i increases, if its fitness exceeds the average fitness of the population. Otherwise it will decline. The total population size remains constant: $\sum_{i=1}^{n} x_i = 1$ and $\sum_{i=1}^{n} \dot{x}_i = 0$.

The set of points with the property $\sum_{i=1}^{n} x_i = 1$ is called the simplex S_n (Figure 2.3). Each point in the simplex refers to a particular structure of the population. The interior of the simplex is the set of points $\vec{x}$ with the property that $x_i > 0$ for all $i = 1, \ldots, n$. The face of the simplex is the set of points $\vec{x}$ with the property that $x_i = 0$ for at least one i. The vertices of the simplex

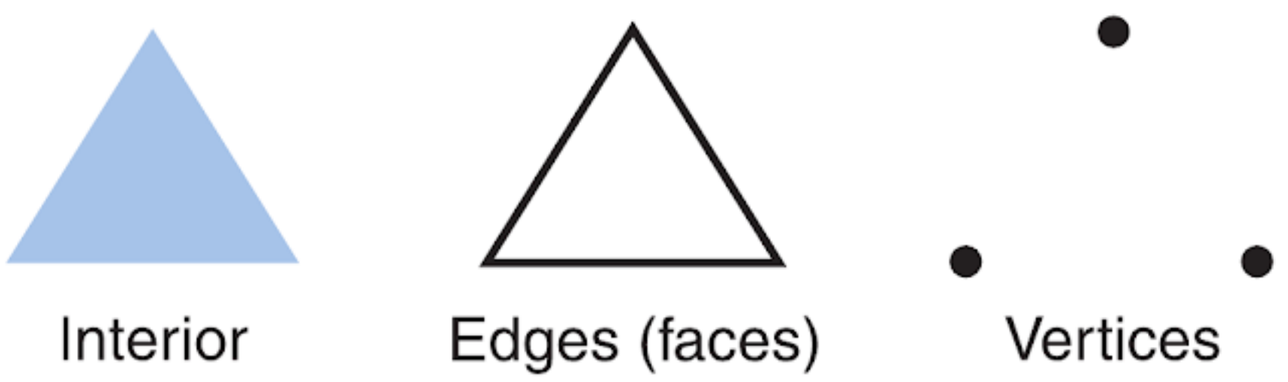


Figure 2.4 The *interior* of a simplex is the set of all points where all coordinates are strictly positive; this means no type has become extinct. The *faces* are the sets of points where at least one coordinate is zero; this means at least one type has become extinct. The *vertices* describe pure populations, where all but one type have become extinct.

are the corner points where exactly one type is present, $x_i = 1$, while all other types are extinct, $x_j = 0$ for all $j \neq i$ (Figures 2.4 and 2.5).

The simplex S_2 is given by the closed interval $[0, 1]$. The notation $[0, 1]$ refers to all numbers which are greater than or equal to 0 and less than or equal to 1. In contrast, $(0, 1)$ is the open interval; it contains all numbers that are strictly greater than 0 and strictly less than 1. The open interval $(0, 1)$ is the interior of the closed interval $[0, 1]$ and, therefore, is also the interior of the simplex S_2.

Equation (2.16) contains a single globally stable equilibrium. Starting from any initial condition in the interior of the simplex, the population will converge to a corner point where all but one type have become extinct. The winner, k, enjoys a well-deserved victory because it has the property of having the largest fitness, f_k. Thus $f_k > f_i$ for all $i \neq k$. The system shows competitive exclusion: the fittest type will outcompete all others. This is the concept of "survival of the fittest."

2.2.2 Survival of the First, Survival of All

Let us return to the selection of two types, A and B, but without making the assumption that their growth rates are linear functions of their frequencies. Instead consider the equation

5 points in S_3

$x_1 = 0$

$x_2 = 0$

$x_1 = 1$

$x_2 = 0$

$x_3 = 0$

$x_1 = 0$

$x_2 = 1$

$x_3 = 0$

Figure 2.5 Five points on the simplex S_3. In the center, $(1/3, 1/3, 1/3)$, all three types have the same frequency. There are three faces. The center of one particular face is given by $(0, 1/2, 1/2)$; one type has become extinct. The corner points (vertices) indicate populations that consist of only one type. S_3 has three corners: $(1, 0, 0)$, $(0, 1, 0)$, and $(0, 0, 1)$.

$$\dot{x} = ax^c - \phi x$$
$$\dot{y} = by^c - \phi y \tag{2.17}$$

As before, a and b denote the fitness values of A and B, respectively. If $c = 1$, we are back to equation (2.13). If $c < 1$, then growth is subexponential. In the absence of the density limitation, ϕ, the growth curve of the two types would be slower than exponential.

In contrast, if $c > 1$, then growth is superexponential. In the absence of the density limitation, ϕ, the growth curve of the two types would be faster than exponential (hyperbolic). To maintain a constant population size, $x + y = 1$, we set $\phi = ax^c + by^c$. Equation (2.17) reduces to

$$\dot{x} = x(1 - x)f(x) \tag{2.18}$$

where

$$f(x) = ax^{c-1} - b(1 - x)^{c-1}. \tag{2.19}$$

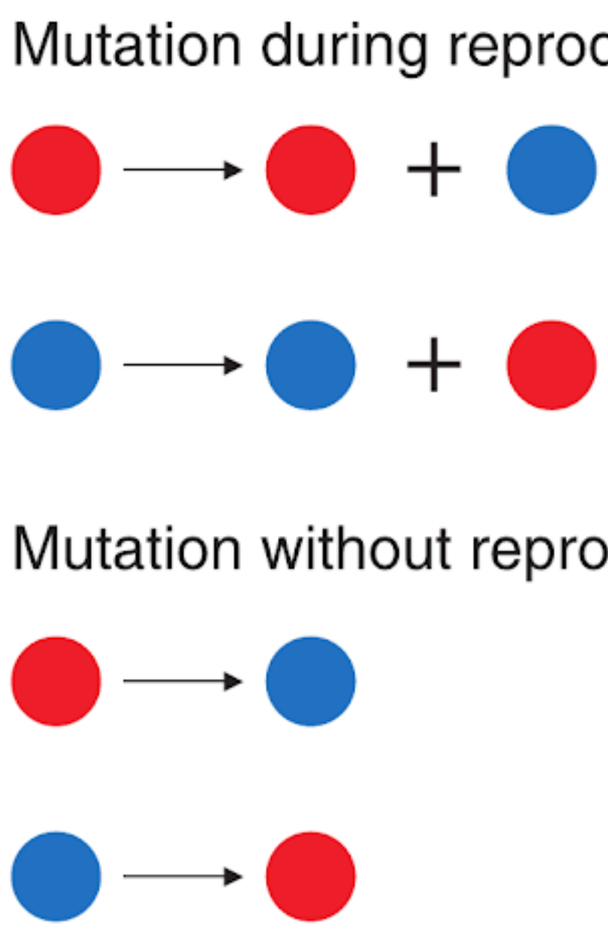


Figure 2.7 Mutation can occur during reproduction: type A produces an offspring that is type B. Mutation can also occur in the absence of reproduction: type A changes into type B. Many genetic mutations occur when the genomic material of a cell is being copied. But mutagens can also change the genetic material of a cell when it is not dividing.

Conversely, denote by u_2 the mutation rate from B to A. As before, let x and y denote the frequencies of A and B, respectively. We have

$$\begin{aligned} \dot{x} &= x(1-u_1) + yu_2 - \phi x \\ \dot{y} &= xu_1 + y(1-u_2) - \phi y \end{aligned} \tag{2.22}$$

Since A and B have the same fitness ($a = b = 1$), the average fitness of the population is constant and given by $\phi = 1$. Taking into account $x + y = 1$, system (2.22) reduces to the differential equation

$$\dot{x} = u_2 - x(u_1 + u_2). \tag{2.23}$$

The frequency of A converges to the stable equilibrium

$$x^* = \frac{u_2}{u_1 + u_2}. \tag{2.24}$$

Hence mutation leads to coexistence between A and B. The relative proportion of A and B at equilibrium depends on the mutation rates. At equilibrium, the ratio of A to B is given by $x^*/y^* = u_2/u_1$. If the mutation rates are the same, $u_1 = u_2$, and then $x^* = y^*$.

Sometimes the mutation rate in one direction is much larger than in the other direction. In these cases, it often makes sense to ignore mutation in the other direction altogether. Let $u_2 = 0$. We have

$$\dot{x} = -x u_1. \tag{2.25}$$

Therefore the frequency of A declines over time as

$$x(t) = x_0 e^{-u_1 t}. \tag{2.26}$$

The frequency of B increases as

$$y(t) = 1 - (1 - y_0) e^{-u_1 t}. \tag{2.27}$$

If mutation occurs only from A to B but not the other way around, then A will die out and B will take over the whole population. We see that mutation can affect survival. Different mutation rates can introduce selection even in the absence of different reproductive rates.

2.3.1 Mutation Matrix

We can extend mutation dynamics to n different types. Let us introduce the mutation matrix, $Q = [q_{ij}]$. The probability that type i mutates to type j is given by q_{ij}. Since each type i has to produce itself or some other type, we have $\sum_{j=1}^{n} q_{ij} = 1$. Thus Q is a stochastic $n \times n$ matrix. A stochastic matrix is defined by the properties that (i) all entries are numbers from the interval $[0, 1]$ (so-called probabilities), (ii) there are as many rows as columns, and (iii) the sum of each row is 1. Stochastic matrices always have 1 as an eigenvalue, and no eigenvalue has an absolute value greater than 1.

Mutation dynamics can be written as

$$\dot{x}_i = \sum_{j=1}^{n} x_j q_{ji} - \phi x_i \qquad i = 1, \ldots, n \tag{2.28}$$

In vector notation we can write

$$\dot{\vec{x}} = \vec{x} Q - \phi \vec{x}. \tag{2.29}$$

Again the average fitness is just $\phi = 1$. The equilibrium is given by the left-hand eigenvector associated with eigenvalue 1:

$$\vec{x}^* Q = \vec{x}^*. \tag{2.30}$$

The point $\vec{x}^*$ denotes the unique globally stable equilibrium of the mutation dynamics.

2.4 MATING

One of the problems that Charles Darwin could not solve was the following: under random mating and blending inheritance, the variability in a population should rapidly decline. Yet it was clear that variability was needed for natural selection. If variability disappears, then natural selection has nothing upon which to act. Suppose there is a distribution of body size in a population. If children inherit the average body size of their parents, then after some time everybody is the same size. Under these circumstances, how can natural selection affect changes in body size?

The first part of the solution is that inheritance (on the level of genes) is not blending but particulate, as had been discovered by Gregor Mendel and published in 1866. That is, individuals have discrete genotypes that get reshuffled, not blended, during mating. Mendel's work was unknown to Darwin. The second step was a simple mathematical analysis, which was performed by the British mathematician G. H. Hardy, who was proud never to have done anything useful (= applied) in his life, only to have his name forever associated with a highly useful and very applied concept in population genetics. Moreover, Hardy's brief calculation was generalized by the German physician Wilhelm Weinberg.

Consider an infinitely large population of a diploid organism with two sexes and random mating (a diploid organism has two copies of its genome; humans and many other animals are diploid). Let us look at one particular gene locus and assume there are two alleles, A_1 and A_2. The alleles are variants of the same gene and might differ in one or a few point mutations. (Point mutation means that only one single base of the DNA sequence is changed.)

There are 3 different genotypes: A_1A_1, A_1A_2, A_2A_2. Let us denote their frequencies in the population by x, y, and z, respectively. Denote by p and q the frequencies of alleles A_1 and A_2. We have $x + y + z = 1$ and $p + q = 1$. Moreover,

$$\begin{aligned} p &= x + \frac{1}{2}y \\ q &= z + \frac{1}{2}y \end{aligned} \tag{2.31}$$

Let us now assume random mating. In the next generation, the genotype frequencies are given by

$$\begin{aligned} x' &= p^2 \\ y' &= 2pq \\ z' &= q^2 \end{aligned} \tag{2.32}$$

For the allele frequencies in the next generation we have again

$$\begin{aligned} p' &= x' + \frac{1}{2}y' \\ q' &= z' + \frac{1}{2}y' \end{aligned} \tag{2.33}$$

Combining (2.32) and (2.33), we observe that

$$p' = p \qquad q' = q \tag{2.34}$$

Therefore the allele frequencies remain unchanged from one generation to the next. Moreover, combining (2.32) and (2.34), we observe

$$\begin{aligned} x' &= p'^2 \\ y' &= 2p'q' \\ z' &= q'^2 \end{aligned} \tag{2.35}$$

From the first generation on, the genotype frequencies can be directly derived from the allele frequencies. Note that equation (2.35) need not hold for the initial genotype and allele frequencies. The Hardy-Weinberg law (expressed by equations 2.34 and 2.35) can be generalized to n alleles.

In summary, the Hardy-Weinberg law states that particulate inheritence preserves variation within a population under random mating.

SUMMARY

- Evolution requires populations of reproducing individuals.
- Asexual reproduction leads to exponential population growth (which will eventually be checked by resource limitation).
- Simple models of population growth in discrete time can give rise to very complicated dynamics.
- Selection arises when different types of individuals reproduce at different rates.
- Normally, the faster-reproducing (fitter) individual outcompetes the slower reproducing (less fit) individual.
- If there are many different types, then selection dynamics can lead to "survival of the fittest." All others become extinct.
- Sublinear growth rates lead to coexistence, "survival of all."
- Superlinear growth rates prevent invasion of a new type and thereby lead to "survival of the first."
- Mutation arises when reproduction is not perfectly accurate.
- Mutation promotes coexistence of different types.
- Asymmetric mutation can lead to selection even if all individuals have the same reproduction rate.
- The Hardy-Weinberg law states that random mating preserves genetic variation within a population.

Figure 3.1 Genomes live in sequence space. The number of dimensions is given by the length of the genome. Small viruses live in 10,000 dimensions. Humans live in about 3 billion dimensions.

in an L-dimensional space. In each dimension there are 4 discrete possibilities. Hence there are 4^L possible sequences.

For writing computer programs, it is often convenient to use binary sequences, the fundamental strings of silicon thoughts. Moreover, everything from Shakespeare to *E. coli* can be encoded in binary sequences. For length L there are 2^L possibilities. In Figure 3.1, the binary sequence space for $L = 3$ is shown. The distance between 000 and 010 is one. The distance between 000 and 011 is 2 (and not $\sqrt{2}$). Hence sequence space is characterized not by a Euclidean metric but by a so-called Hamming metric or Manhattan metric. In Manhattan, if you are on 5th Avenue and 51st Street it takes 2 blocks to go to 6th Avenue and 52nd Street, not $\sqrt{2}$ blocks. This metric was introduced by Richard Hamming in information theory.

Let us compare the binary sequence space of length $L = 300$ with a three-dimensional cubic lattice containing the same number of points. There are

Fitness landscape = each sequence has a reproduction rate (= fitness)

Figure 3.2 The fitness landscape is a high-dimensional mountain range. Each genome (= each point in sequence space) gets assigned a fitness value.

Sequence space

$2^{300} \approx 10^{90}$ points. Imagine nearest neighbors are placed at a distance of 1 meter. The diagonal of the three dimensional cubic lattice has a length of about 10^{30} meters, which corresponds to about 10^{14} light years. In contrast, the longest distance in the L-dimensional hypercube is only 300 meters. Thus sequence space is characterized by short distances, but many dimensions. It is not far to move from one sequence to another, but there are many possible steps that lead in wrong directions. Evolution is a trajectory through sequence space. This trajectory needs an efficient guide.

3.2 FITNESS LANDSCAPES

The American population geneticist Sewall Wright invented the concept of a "fitness landscape" in the 1930s, but Manfred Eigen and Peter Schuster, collaborating in the 1970s, combined fitness landscape with sequence space. Consider a function that assigns to each genomic sequence a fitness value. Hence we build a mountain range on the foundation of an L-dimensional sequence space (Figure 3.2). This mountain range has $L + 1$ dimensions. The evolutionary process of mutation and selection explores this hyper-alpine mountain range.

The genomic sequence represents the genotype of an organism. The phenotype of an organism is given by its shape, behavior, performance and any kind of ecological interaction. The phenotype determines the fitness (reproductive rate) of the organism. There is a mapping from genotype to phenotype.

A quasispecies is a population of reproducing RNA or DNA molecules

4-nucleotide alphabet Binary alphabet

Figure 3.3 The ensemble of genomes of a natural population form a quasispecies: the genomes of different individuals are similar but not identical. Biology has chosen a four-letter alphabet consisting of the nucleotides A, T, C, and G for its genes. Most in silico evolution uses a binary alphabet for convenience. Sequence differences (mutations) are shown in red.

There is another mapping from phenotype to fitness. The fitness landscape is a convolution of these two mappings. It is a direct mapping from genotype to fitness.

The fitness landscape of certain problems can be determined experimentally. For example, HIV can generate point mutations that confer drug resistance. The relative growth rate of such mutants can be determined by in-vitro assays. In general, however, to understand the relationship between genotype, phenotype, and fitness is an extremely complicated problem. Much of biology, including developmental biology, molecular biology, post-genomics, and proteomics, is devoted to this very task.

3.3 THE QUASISPECIES EQUATION

A quasispecies is an ensemble of similar genomic sequences generated by a mutation-selection process (Figure 3.3). The term was introduced by the chemists Manfred Eigen and Peter Schuster. In chemistry the word "species" refers to an ensemble of identical molecules, for example, the species of all water molecules. But the species of all RNA molecules does not contain identical sequences, and therefore the term "quasispecies" was coined. Biologists

are sometimes confused by this expression, because they relate it to the concept of a biological species.

We stay with binary sequences for convenience. We note that any genomic or other information can be encoded by binary sequences. Consider all binary sequences of length L. Enumerate all those sequences by $i = 0, 1, 2, \ldots, n$ where $n = 2^L - 1$. A natural enumeration is obtained if the sequence represents the binary description of the corresponding integer. For example, let $L = 4$. The sequence 0000 corresponds to $i = 0$, the sequence 0001 to $i = 1$, the sequence 0010 to $i = 2, \ldots,$ the sequence 1111 to $i = 15$.

Imagine an infinitely large population of organisms, each carrying a genome of length L. Denote by x_i the relative abundance (= frequency) of those organisms that contain genome i. We have $\sum_{i=0}^{n} x_i = 1$. The genomic structure of the population is given by the vector $\vec{x} = (x_0, x_1, \ldots, x_n)$.

Denote by f_i the fitness of genome i. It is a non-negative real number. Thus genomes of type i are being reproduced at rate f_i. The fitness landscape is given by the vector $\vec{f} = (f_0, f_1, \ldots, f_n)$. The average fitness of the population, $\phi = \sum_{i=0}^{n} x_i f_i$, is the inner product of the vectors $\vec{x}$ and $\vec{f}$. We have $\phi = \vec{x}\,\vec{f}$.

During replication of a genome, mistakes can happen. The probability that replication of genome i results in genome j is given by q_{ij}. Here we again meet the mutation matrix $Q = [q_{ij}]$ of section 2.3. We remember that Q is a stochastic matrix: it has as many rows as columns; each entry is a probability, which means a number between 0 and 1; each row sums to one, $\sum_{j=0}^{n} q_{ij} = 1$.

The quasispecies equation (Figure 3.4) is given by

$$\dot{x}_i = \sum_{j=0}^{n} x_j f_j q_{ji} - \phi x_i \qquad i = 0, \ldots, n \tag{3.1}$$

Sequence i is obtained by replicating any sequence j at rate f_j times the probability that replication of sequence j generates sequence i. Each sequence is removed at rate ϕ to ensure that the total population size remains constant, $\sum_{i=0}^{n} x_i = 1$. Thus quasispecies dynamics are defined on the simplex, S_n.

In the limiting case of completely error-free replication, Q becomes the identity matrix: all diagonal entries are one, all off-diagonal entries are zero.

The quasispecies equation

Figure 3.4 The quasispecies equation, formulated by Manfred Eigen and Peter Schuster, is one of the most important equations in theoretical biology. It describes the mutation and selection of an infinitely large population on a constant fitness landscape.

Consider an initial condition in the interior of the simplex, defined by $x_i > 0$ for all i. The quasispecies will converge to a homogeneous population that consists only of the fittest sequence. If $f_0 > f_i$ for all $i \neq 0$, then the stable equilibrium is given by $x_0 = 1$ and $x_i = 0$ for $i \neq 0$. If there are no errors, then the quasispecies equation (3.1) reduces to the selection equation (2.16) of section 2.2.1.

Let us now assume that errors occur. This means that (at least some) off-diagonal entries of Q are not zero. In many realistic contexts, the matrix Q is irreducible, which means it is possible to find a sequence of mutations from any one genome i to any other genome j. Furthermore, let $f_i > 0$ for at least some i. In this case, the quasispecies equation admits a single, globally stable equilibrium, $\vec{x}^*$, in the simplex S_n.

The equilibrium quasispecies, $\vec{x}^*$, does not necessarily maximize the average fitness ϕ. Consider again a fitness landscape with the property $f_0 > f_i$ for all $i \neq 0$. Then the population consisting only of sequence 0 will have a higher fitness than the equilibrium population $\vec{x}^*$. Thus, mutations reduce the average fitness at equilibrium.

Observe that (3.1) is a nonlinear differential equation. The term $-\phi x_i$ is of second order. Linear differential equations can always be solved, but nonlinear differential equations normally cannot be solved. This means for nonlinear differential equations the trajectories cannot always be written as explicit

Hence a mutation has to occur in as many positions as differ between the sequences i and j, which is precisely the Hamming distance, h_{ij}. No mutation must occur in the remaining $L - h_{ij}$ positions.

Equation (3.11) is an elegant description of a mutation matrix that allows point mutations among binary sequences of constant length. It is assumed that the point mutation rate, u, is the same for all positions. It is further assumed that a mutation in one position is independent of a mutation in another position. Hence one error does not increase the probability of another error. There are no insertions and no deletions. All of these restrictions can be relaxed in principle, but doing so will lead to considerable complexity.

Let us use mutation matrix (3.11) to describe the human immunodeficiency virus as an example. The point mutation rate of HIV is approximately $u = 3 \times 10^{-5}$. The genome length of HIV is $L = 10^4$. Therefore the probability that the whole HIV genome is replicated without mutation is given by $(1 - u)^L \approx 0.74$. The probability that replication of the HIV genome results in a sequence that differs in one arbitrary position is given by $Lu(1 - u)^{L-1} = 0.22$. The probability that a particular one-error mutant, for example one that confers drug resistance or immune escape, is being produced is given by $u(1 - u)^{L-1} = 2.2 \times 10^{-5}$. If 10^9 newly infected cells are being produced each day, then any particular one-error mutant will arise 22,000 times each day. This number signifies the enormous potential of HIV (or other viruses or microbes) to escape from selection pressures that are meant to control them. We will revisit this topic in Chapter 10.

3.5 ADAPTATION IS LOCALIZATION IN SEQUENCE SPACE

The quasispecies equation (3.1) describes the movement of a population through sequence space. The quasispecies "feels" gradients in the mountain range of the fitness landscape. It attempts to climb uphill and reach local or global peaks (Figure 3.5). What are the conditions that this evolutionary walk will be successful? One such condition is the error threshold.

If the mutation rate u is too high, then the ability of the quasispecies to climb uphill and to remain on top of a mountain peak is impaired. In fact, we can show that for many natural fitness landscapes there is a maximum mutation rate, u_c, that is still compatible with adaptation. If the mutation rate exceeds this value, $u > u_c$, then adaptation is not possible.

Evolution is adaptation of the quasispecies on the fitness landscape

Figure 3.5 Quasispecies love to climb mountains in high-dimensional spaces. The higher they get, the fitter they are. Adaptation means to go up.

Sequence space

Adaptation means that the quasispecies is able to find peaks in the fitness landscape and stay there. Suppose the fitness landscape contains only one peak. If the mutation rate is sufficiently low, then the equilibrium solution of equation (3.1) describes a quasispecies that is centered on this peak. Most sequences resemble the type with maximum fitness or nearby mutants. Sequences that are far away from the peak will have a very low frequency. (In population genetics, frequency means relative abundance.) We say the quasispecies is adapted to this peak. Similarly, we can say that the quasispecies distribution is localized at this peak. Adaptation means localization in sequence space. When the mutation rate of a quasispecies is zero, it contains only sequences with maximum fitness. When the mutation rate is very small, the quasispecies distribution is very narrow. As the mutation rate increases, the quasispecies distribution widens. There is a critical mutation rate, u_c, beyond which the equilibrium quasispecies no longer "feels" the peak. The quasispecies is no longer localized around the peak. Adaptation is lost. Strictly speaking, a well-defined "phase transition" from a localized to a delocalized state only occurs for infinite sequence length, but the phenomenon is striking already for binary sequences of length $L = 10$.

The maximum mutation rate, u_c, that is compatible with adaptation is called the "error threshold." Not all fitness landscapes have error thresholds. Narrow peaks of finite height have error thresholds. If a peak is so broad that most sequences in the sequence space are within the slopes of the peak, then an error threshold need not occur.

Quasispecies have a tendency to climb uphill. Starting from some random initial condition, $\vec{x}(0)$, the quasispecies equation (3.1) will tend to increase the average fitness, ϕ. But it is also easy to construct a counterexample. Suppose a certain sequence has maximum fitness, while all other sequences have lower fitness. If we start with a population that contains only the sequence with maximum fitness, then equation (3.1) will reduce the average fitness ϕ until an equilibrium between mutation and selection, a so-called mutation-selection balance has been reached.

Calculating the error threshold, u_c, for complex fitness landscapes is difficult, but the following simple fitness landscape provides the crucial insight. Consider all binary sequences of length L. The all-zero sequence, $00\ldots0$, has the highest fitness given by $f_0 > 1$. All other sequences have fitness 1. The all-zero sequence is sometimes called the "master sequence" or the wild type, while all other sequences are called "mutants."

The probability that the master sequence produces an exact copy of itself is given by $q = (1-u)^L$. The probability that the master sequence generates any mutant is given by $1-q$. The trick is to neglect the back mutation from the mutants to the master sequence. With this assumption the quasispecies equation (3.1) becomes

$$\begin{aligned} \dot{x}_0 &= x_0(f_0 q - \phi) \\ \dot{x}_1 &= x_0 f_0(1-q) + x_1 - \phi x_1 \end{aligned} \tag{3.12}$$

Here x_0 is the frequency of the master sequence, while x_1 is the sum of the frequencies of all the mutants. Clearly, $x_0 + x_1 = 1$. The average fitness is given by $\phi = f_0 x_0 + x_1$. System (3.8) collapses to a single equation

$$\dot{x}_0 = x_0[f_0 q - 1 - x_0(f_0 - 1)]. \tag{3.13}$$

If $f_0 q < 1$, then x_0 will converge to zero; the fittest sequence cannot be maintained in the population. If $f_0 q > 1$, then x_0 will converge to

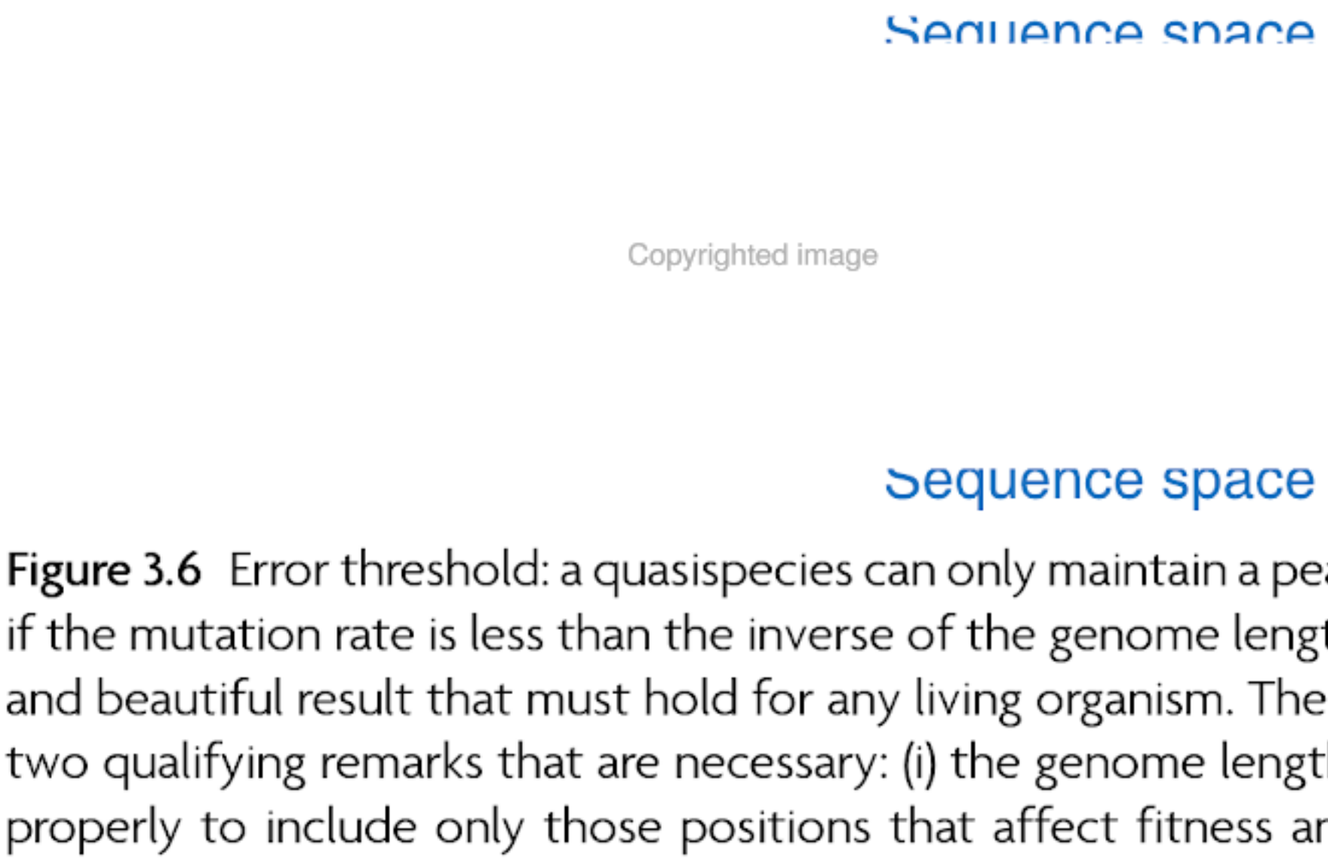
$$x_0^* = \frac{f_0 q - 1}{f_0 - 1}. \tag{3.14}$$

Hence, the error threshold is given by

$$f_0 q > 1. \tag{3.15}$$

Error threshold: adaptation is only possible if the mutation rate per base, u, is less than the inverse of the genome length, L

$$u < 1/L$$



$$u > 1/L$$

Figure 3.6 Error threshold: a quasispecies can only maintain a peak in a fitness landscape if the mutation rate is less than the inverse of the genome length. This is a very general and beautiful result that must hold for any living organism. The beauty is not spoilt by two qualifying remarks that are necessary: (i) the genome length, L, has to be defined properly to include only those positions that affect fitness and (ii) there are some pathological landscapes where a peak can be maintained beyond the error threshold, for example if the peak is "infinitely" high or so wide that its presence can be felt by the majority of all possible sequences.

This inequality can be rewritten as $\log f_0 > -L \log(1-u)$. For small mutation rates, $u \ll 1$, we have $\log(1-u) \approx -u$. Therefore we obtain the condition

$$u < \frac{\log f_0}{L}. \tag{3.16}$$

If the fitness advantage of the master sequence is not too large and not too small, then $\log f_0$ is approximately 1. Now the error-threshold condition reduces to

$$u < 1/L. \tag{3.17}$$

Hence the maximum mutation rate that is still compatible with adaptation has to be less than the inverse of the genome length (Figure 3.6). In other

Table 3.1 Genome length (in bases), mutation rate per base, and mutation rate per genome for organisms ranging from DNA viruses to humans

Organism	Genome length in bases	Mutation rate per base	Mutation rate per genome
RNA viruses			
Lytic viruses			
Qβ	4.2×10^3	1.5×10^{-3}	6.5
Polio	7.4×10^3	1.1×10^{-4}	0.84
VSV	1.1×10^4	3.2×10^{-4}	3.5
Flu A	1.4×10^4	7.3×10^{-6}	0.99
Retroviruses			
SNV	7.8×10^3	2.0×10^{-5}	0.16
MuLV	8.3×10^3	3.5×10^{-6}	0.029
RSV	9.3×10^3	4.6×10^{-5}	0.43
Bacteriophages			
M13	6.4×10^3	7.2×10^{-7}	0.0046
λ	4.9×10^4	7.7×10^{-8}	0.0038
T2 and T4	1.7×10^5	2.4×10^{-8}	0.0040
E. coli	4.6×10^6	5.4×10^{-10}	0.0025
Yeast *(S. cerevisiae)*	1.2×10^7	2.2×10^{-10}	0.0027
Drosophila	1.7×10^8	3.4×10^{-10}	0.058
Mouse	2.7×10^9	1.8×10^{-10}	0.49
Human *(H. sapiens)*	3.5×10^9	5.0×10^{-11}	0.16

Sources: Drake (1991, 1993) and Drake et al. (1998).

Note: Most organisms have a mutation rate per genome which is less than one, as predicted by the error threshold theory. Why Qβ and VSV have such a high mutation rate is at present unexplained.

words, the genomic mutation rate, uL, has to be less than one. In fact, this condition holds for most living organisms for which mutation rates have been measured (Table 3.1). For eukaryotes, the genome length L in this context should actually be defined as the total number of bases in the coding and regulatory regions of the DNA.

3.6 SELECTION OF THE QUASISPECIES

The following remarkable observation was first made by Peter Schuster and Jörg Swetina. Consider a fitness landscape that contains a high but narrow

- Adaptation is localization in sequence space. This is only possible if the mutation rate is below the error threshold.
- The error threshold states that the maximum possible mutation rate (per base) must be less than the inverse of the genome length (in bases).

EVOLUTIONARY GAMES

EVOLUTIONARY GAME THEORY means that the fitness of individuals is not constant, but depends on the relative proportions (frequencies) of the different phenotypes in the population: fitness is frequency dependent. Evolutionary game theory is the generic approach to evolutionary dynamics and contains as a special case constant selection.

Game theory was invented by John von Neumann and Oskar Morgenstern. They wanted to design a mathematical theory to study human behavior in strategic and economic decisions. Von Neumann was a Hungarian-born mathematician working at the Institute for Advanced Study, where he invented and revolutionized several fields of mathematics. We have already encountered him in Chapter 3 in connection with the terms "translation" and "transcription," which he invented when thinking about how to conceive of a machine that could reproduce itself. He built the first computer that held the program for the calculation in its memory rather than in its hardware. Incidentally, one of the first projects that this computer did in its spare time was a mathematical simulation of an evolutionary system.

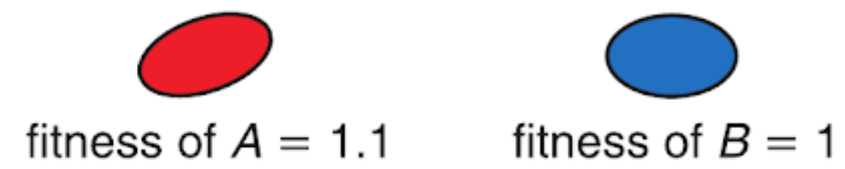


Figure 4.1 Constant selection means fitness neither depends on the composition of the population nor changes over time. For example, A has constant fitness 1.1, while B has constant fitness 1. In contrast, frequency-dependent selection means that fitness does depend on the relative abundance (= frequency) of individual types. A has the ability to move. If few other cells are moving, then A has a larger fitness than B. But if many other cells "are on the road," this fitness advantage is reversed (in this hypothetical example).

of the frequency of A. A has a higher fitness than B when A is rare, but has a lower fitness than B when A is common. What is the outcome of such a selection process?

Let us formalize the general case of frequency-dependent selection between two strategies A and B. Denote by x_A the frequency of A and by x_B the frequency of B. The vector $\vec{x} = (x_A, x_B)$ defines the composition of the population. Denote by $f_A(\vec{x})$ the fitness of A and by $f_B(\vec{x})$ the fitness of B. The selection dynamics can be written as

$$\begin{aligned} \dot{x}_A &= x_A[f_A(\vec{x}) - \phi] \\ \dot{x}_B &= x_B[f_B(\vec{x}) - \phi] \end{aligned} \tag{4.1}$$

The average fitness is given by $\phi = x_A f_A(\vec{x}) + x_B f_B(\vec{x})$.

Because $x_A + x_B = 1$ at all times, we can introduce the variable x with $x_A = x$ and $x_B = 1 - x$. We can write the fitness functions as $f_A(x)$ and $f_B(x)$.