MICROBIOLOGY AND MOLECULAR BIOLOGY REVIEWS, Mar. 2001, p. 151–185 Vol. 65, No. 1
1092-2172/01/$04.00+0 DOI: 10.1128/MMBR.65.1.151–185.2001

Transition between Stochastic Evolution and Deterministic Evolution in the Presence of Selection: General Theory and Application to Virology

I. M. ROUZINE,[1]* A. RODRIGO,[2] AND J. M. COFFIN[1]

Department of Molecular Biology and Microbiology, Tufts University, Boston, Massachusetts 02111,[1] and School of Biological Sciences, University of Auckland, Auckland, New Zealand[2]

* Corresponding author. Mailing address: Department of Molecular Biology and Microbiology, Tufts University, Boston, MA 02111. Phone: (617) 782-3872. Fax: (617) 636-4086. E-mail: irouzine@emerald.tufts.edu.

INTRODUCTION

The process of evolution is a consequence of the interplay of mutation, selection, and chance on a population of organisms, leading to an observable change in its genetic makeup. Since the time of Darwin, the influence of these factors on the evolution of organisms ranging from bacteria to humans has been intensively studied, both experimentally and theoretically, leading to a very large body of literature. Only recently, however, has attention been turned toward special problems in the evolution of viruses. Virus evolution is of particular interest and importance for three reasons. First, we desire to gain an understanding (usually in the absence of a fossil record) of how modern viruses have arisen from their earlier forms, both in recent times and in parallel with the evolution of their hosts. Second, the evolution of a virus during the course of infection of a single host, or along a short transmission chain, is of great importance in creating new populations with properties altered in important ways, such as evasion of the immune response, resistance to antiviral therapy, or altered virulence. Third, because of their high replication rates, simple genomes, large population sizes, and high mutation rates, viruses make good models for studying and testing evolutionary theory.

Particular attention has focussed on understanding the evolutionary forces that act on human immunodeficiency virus (HIV) during the course of infection of a single human host. HIV displays a remarkable extent of genetic variation concurrent with a high speed of evolution: in the most variable region of the genome (*env*), individual genomes within a population from an infected person can vary by as much as 3 to 5% (2, 43, 78); substitutions in *env* accumulate at a rate of approximately 1% per year (71), 50 million times faster than in the small subunit of rRNA (61). This variation has important consequences. It allows the virus to evolve to infect different cell types (9, 20, 30) and to rapidly become resistant to otherwise highly effective antiviral drugs (10, 47, 50); it may play a role in evading the immune system (4, 56, 73, 79). Furthermore, its high mutation rate (estimated to average about 3×10^{-5} per nucleotide site per replication cycle [49]), large population size (variously estimated from about 10^7 to 10^8 productively infected cells), and continuous steady state, in which the large majority of virions and productively infected cells turns over every day (25, 77), create a situation which, at least in principle, is amenable to (and requires) mathematical modeling.

To date, a number of modeling approaches have been applied to understand the evolution of HIV in vivo. These approaches use either population genetic (mutation frequency distribution) or phylogenetic inference using virus sequences obtained from HIV-infected individuals. In general, they are based on one of two different theoretical frameworks to the evolution problem. Deterministic approaches, including quasispecies theory (15, 26), assume that the population size is very large, such that the frequency of a given mutation at any given time is completely predictable if one knows the initial frequency, the mutation rate, and the selection coefficient (i.e., the differential growth rate conferred by the different alleles). At first glance, such approaches would seem justified by the large number of infected cells at each generation (21); however, a number of factors, such as variation in the replication potential and generation times among infected cells, may lead to an effective population size much smaller than the actual number of infected cells. Stochastic models, as applied to HIV (to this point), proceed from the opposite assumption: that the effective population size is so small (or that selective forces are so weak) that random drift dominates over selection. The hypothesis of selectively neutral mutations has a long, successful history in describing the evolution of organisms where populations are small (and not uniformly distributed) and mutation rates are very low (36). Their applicability to virus populations remains to be established. Many of the assumptions that underlie neutral theory are not appropriate for virus populations, and a number of characteristics of HIV genetic variation in vivo, such as the uneven ratio of synonymous to nonsynonymous changes in different regions of the genome (5, 44, 48), argue against simple application of neutral theory. However, inclusion of selection effects in evolutionary analysis (for example, the coalescent method) presents a mathematical challenge that has not yet been fully solved in a practical fashion, although progress toward this goal has been made recently (42, 55).

As an example of the difference between deterministic and stochastic models, consider the question of the frequency in a population of a mutation that is slightly deleterious to virus replication. In a deterministic system, it can be easily calculated that the frequency of such a mutation in the population will come to equilibrium at a point equal to the mutation rate divided by the selection coefficient (24). In a stochastic system, the population will usually be completely uniform in one variant or the other (76), switching rarely but rapidly from one

form to the other. This theoretical experiment is of great practical importance in that it describes the appearance of a mutation that can confer resistance to an antiviral drug even before treatment.

To solve this problem and many others, it is clear that a more general theoretical framework is needed: one that takes into account both selection and drift under a set of assumptions more appropriate to viruses than is found in theoretical works published to date. Our aim in this work was to develop, from first principles, a general theory that includes the effects of both selection and drift on a population. We use a set of assumptions appropriate to virus populations, focusing on the interplay between deterministic and stochastic behavior in the context of virologically realistic experiments. We apply these to the simplest possible model: mutation at a single site with only two alleles, replicating in a steady-state system (that is, a constant number of infected cells) under the influence of constant selective pressure in a single isolated population. Because we are dealing with a single locus, we do not consider recombination explicitly; because we are dealing with haploid populations, we do not have to consider allelic dominance. It should be noted that although we do not consider recombination explicitly, the presence of strong recombination must be, in fact, implied for the one-locus approximation to be quantitatively correct. Also, nonconserved loci must be spaced sufficiently far apart in the genome, depending on the recombination rate. Even in the absence of recombination, the one-locus approximation is a useful starting point for understanding interactions between selection and stochastic factors at a qualitative level. We present a complete model that considers the full range of possible values for population size, mutation rate, and selection effects. Despite its simplicity, the model is surprisingly rich in its descriptive power. At the extremes, the results of this model correspond to the standard results of deterministic or neutral theory; however, we have found that there is a large range of values for the key parameters in which the system behaves in an intermediate fashion: under some conditions its evolution is dominated by stochastic factors, whereas at other times it behaves in a nearly deterministic fashion. We refer to this range of parameter values as the "selection-drift" regime and describe its properties in detail.

This work is divided into two major parts. In the first, we present all the principal results in qualitative terms, using language appropriate for a reader trained in biology and with a moderate level of mathematical sophistication. This part is accompanied by a number of illustrative examples obtained by computer simulation. Although keyed to the mathematical formalism of the second part, it is designed to be read independently and to provide the reader with an understanding of the principal results and their biological significance, particularly in the context of virus populations. The second part is a formal mathematical derivation of the principal results of the model. These results are listed at the beginning of each section and derived in the following subsections. Although some of the derivation presented is not novel, in that it parallels classic work of a number of population biologists (18, 19, 23, 24, 31, 37, 81, 82), its formal application specifically to virus systems is, to the best of our knowledge, a new approach, and we present it in full for this reason, as well as to provide a thorough and self-contained review. Although some of our mathematical methods differ from the classic methods, the final results are identical.

The presentation in both parts of this work proceeds in parallel. We first develop the basic evolution equation, which describes, at least in a statistical sense, the change in frequency of a mutant allele as a function of time and the key parameters: mutation rate, selection coefficient, and population size. We then present the predicted results, for all three regimes, of a set of virological experiments: accumulation and reversion of deleterious mutations, competition between mutant and wild-type viruses, gene fixation, mutation frequencies at the steady state, divergence of two populations split from one population, and genetic turnover within a single population. Next, we discuss sampling statistics and the application of this theory to some specific real-world experimental issues of virus and organismal evolution. Finally, we discuss the application and extension of this theoretical framework to other problems, including multilocus evolution and phylogenetic analysis.

QUALITATIVE DISCUSSION AND COMPUTER SIMULATIONS

Description of the Model and the Evolution Equation

In this section, we introduce the population model and explain how to approach the problem of evolution when random factors enter the picture. First we describe a one-locus, two-allele population model based on the virus replication cycle and discuss briefly the main factors of evolution included in the model. This is followed by a discussion of the biological meaning of the evolution equation. Finally, the boundary conditions for the evolution equation describing the properties of a weakly polymorphic population are described.

Virus population model. First, we choose a basic model of virus evolution. For the purposes of simplicity, we consider the evolution of one nucleotide position at a time, and we assume that each nucleotide has a choice between only two alleles. (Such a model applies directly to multiple loci if the evolving loci are sufficiently distant and the recombination rate is sufficiently high. Evolution at closely situated loci or in the absence of efficient recombination is not independent [see "Many loci and other aspects" below].) Conventionally, we denote the better-fit allele as wild type and the less-fit allele as mutant. A deleterious mutation event (from wild type to mutant) will be referred to as forward mutation, and an advantageous mutation event will be referred to as reverse mutation. Each separate nucleotide will be characterized by two parameters, both of which are assumed to be much less than unity: the mutation cost (or selection coefficient), s, which is the relative difference in fitness between the two alleles, and the mutation rate per base per replication cycle, μ. We assume that mutations at different nucleotides have a weak additive effect on virus fitness. In doing so, we neglect epistasis (coselection) arising due to biological interaction between nucleotides at both the nucleotide and protein levels. We also ignore linkage disequilibrium between loci due to random drift, so that different nucleotides evolve independently (see the Introduction). The mutation rate is set, in our work, to be the same in the forward and reverse directions. For example, for HIV in infected cells the mutation rate per base is in the range of 5×10^{-6} to $5 \times$

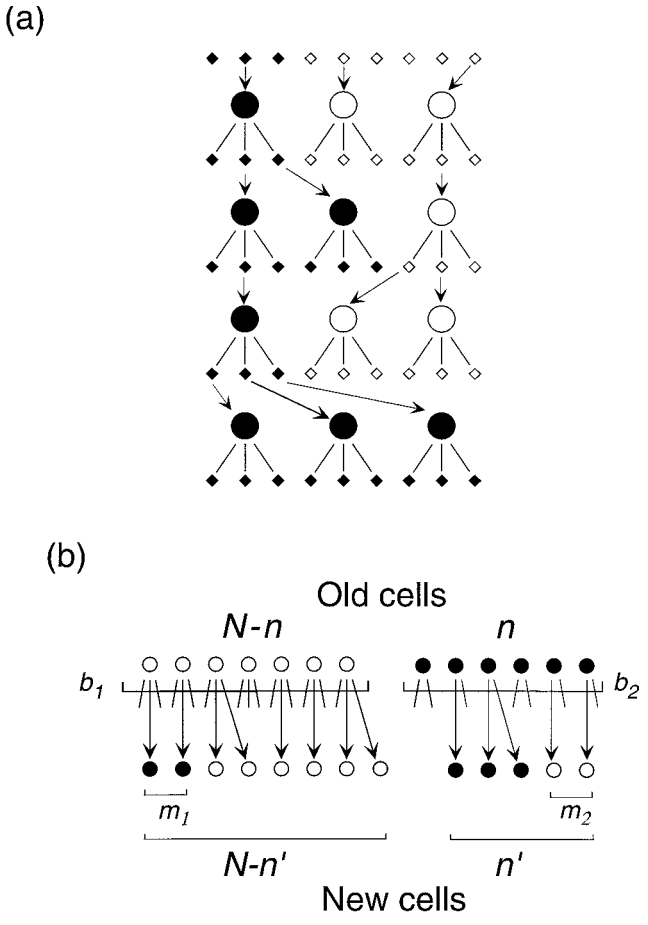

FIG. 1. (a) Drift of genetic composition due to random sampling of infecting virions. Circles denote infected cells, and small diamonds show free virus particles. Black and white denote virus genetic variants. (b) Full virus population model including random drift, selection, and mutation. Two consecutive generations of infected cells are shown. Lines radiating from a cell denote virions, some of which, as shown by arrows, infect new cells. Mutant cells yield fewer progeny per cell. A small fraction of infecting virions, m_1 and m_2, mutate to the other variant.

10^{-5}, depending on the type of substitution (49, 68). The selection coefficient will vary over a wide range according to the specific base and to the specific conditions of replication, but it is assumed to be constant over the period of observation; in other words, there is no selection for diversity.

The basic model of virus replication is illustrated in Fig. 1. Consider the dynamics of a cell population infected by two genetic variants of a virus: a fraction (f) of cells is infected by the mutant virus, and the remaining cells ($1 - f$) are infected by the wild type. The number of mutant-infected cells may change with time, i.e., with each new generation of cells. The total cell count is assumed to be constant. During a generation step, each cell produces a fixed (large) number of virions and then dies and is replaced by an uninfected cell. The number of virions produced and capable of infecting new cells differs, by a factor of $1 - s$, between cells infected with different variants, creating selection for the better-fit (more prolific) variant. Since the total number of infected cells is fixed and the number of virions produced per cell is large, only a small fraction of the virions infect the next generation of cells. On infecting a cell, each virion has a small chance of mutating into the opposite genetic variant, given by the mutation rate introduced above. All the virions produced by a cell afterwards represent the same genetic variant. Thus, intracellular interference between variants does not occur. (This lack of intracellular competition is a reasonable assumption for retroviruses or when the proportion of infected cells in a tissue is much lower than 100%. It may vary in other virus models, when the multiplicity of infection is high.)

Some details of the model, such as fixed burst sizes and the point of the replication cycle at which mutation occurs, are of no consequence when long timescales are considered. Overlap in time between generations of infected cells was neglected but causes a factor of 2 increase in the rate of random drift (52). By contrast, such assumptions as two variants per base and the absence of both coselection and selection for diversity are essential. The model includes a minimal set of three factors of genetic evolution: random drift due to sampling of genomes, mutation, and selection. Let us characterize briefly the effect of each of these factors on the composition of the population as it changes with time.

The model assumes that the virions infecting each new generation of cells are chosen randomly from the virions produced by the mutant and wild-type subpopulations. As a result of this random sampling of genomes, the mutant frequency experiences random drift in time (18, 80), as shown in Fig. 1a. In the absence of mutation and selection, any population composed originally of a mixture of alleles eventually becomes uniform in either genotype (i.e., the allele is fixed), with the probabilities depending on the initial composition.

Selection enters our model through the difference in the number of infectious progeny produced by cells infected with different genetic variants. Selection alone drives the system into a state consisting entirely of the better-fit variant.

Mutations, in contrast to random drift and selection, favor inhomogeneity. If the other two factors are absent, mutations push the system toward the equilibrium composition at which the total numbers of forward and reverse mutations per generation are in balance. For equal forward and reverse mutation rates assumed here, equilibrium occurs at 50% of each allele.

If all three factors are at work and there are no external perturbations, the population will eventually reach a dynamic steady state in which mutation, on average, is in balance with selection and/or random drift. In the steady state, the statistical properties of the population no longer vary with time; i.e., even though the genetic composition may fluctuate strongly with time, all the mean values, standard deviations, etc., remain constant. The whole model with the three factors of evolution is illustrated in Fig. 1b.

Stochastic equation of evolution. Different meanings can be assigned to the word "evolution." For the task at hand, evolution of the population is characterized by the dependence of the frequency of cells infected with mutant virus on time. In deterministic dynamics, which applies only in very large populations of infected cells, if one knows the initial mutant frequency and has the appropriate equations, one can, in principle, predict the mutant frequency at later times with arbitrary precision. (In practise, the equations are never known exactly, since there are many different factors in play, but this is a separate

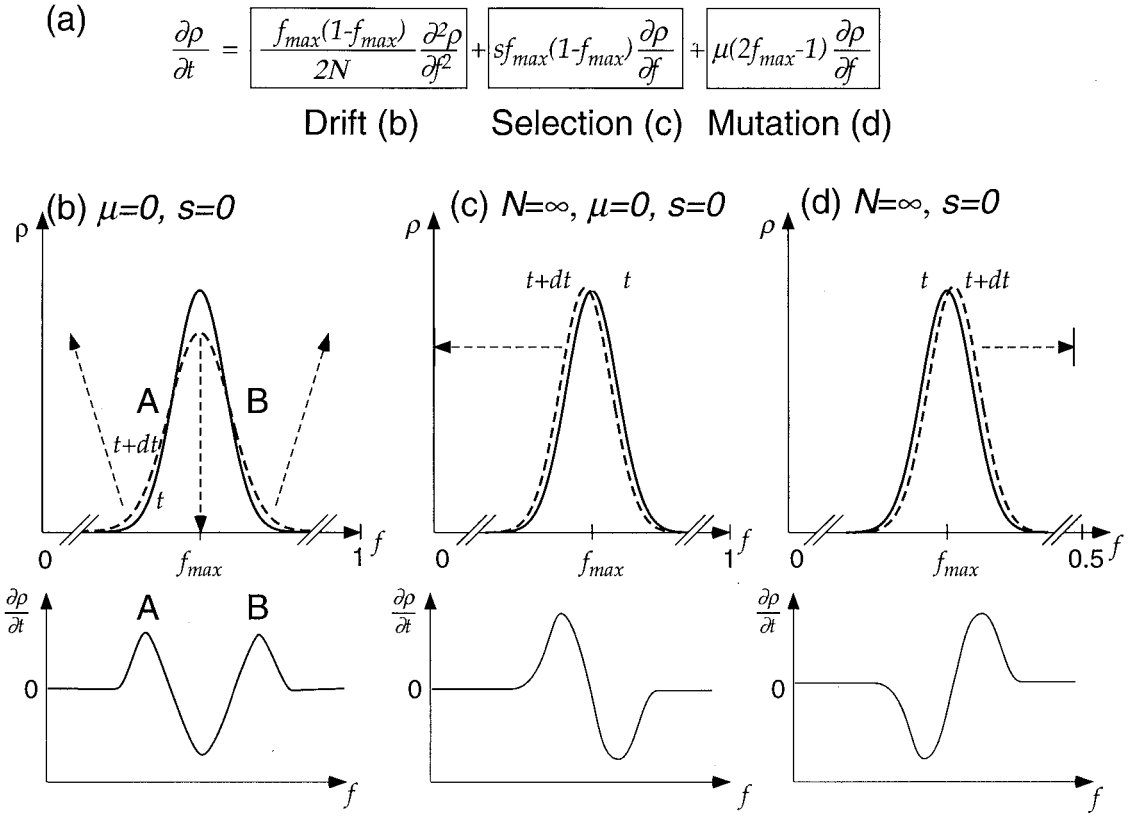

FIG. 2. Illustration of the stochastic evolution equation. (a) The equation shown is derived from equations 1 and 2 in the particular case where the probability density, $\rho(f,t)$, is a narrow peak. Its right-hand side is a composite of three terms describing the effects of random drift, selection, and mutation on the change in the probability density with time. (b to d) The lower panels show the local changes in $\rho(f,t)$, corresponding to each part of the right-hand side of equation in panel a, and the upper panels show the resulting effects: spread (b) and shift (c and d) of the peak. Solid and dashed lines show the peak at two adjacent moments in time.

issue [68].) By contrast, in the presence of random factors, the time dependence of the mutant fraction cannot be predicted even in principle. Even if one knows its precise initial value, the error with which one can predict its value later grows with time. If random factors are strong, the error in the mutant frequency and its value become eventually comparable. Evolution of the mutant frequency, in other words, is a random process.

Randomness of mutations does not mean, however, that the evolution of a population is totally arbitrary. On the contrary, useful predictions can be made about its statistical properties even if its specific state cannot be predicted. Instead of time dependence of the mutant frequency, one has to consider the time-dependent probability density [$\rho(f)$], defined as the chance that a given population has a mutant frequency near a particular value. The probability density, which can be introduced if both subpopulations (mutant and wild type) are large, is closely related to a histogram derived by plotting the number of times the mutant frequency of a population is observed to lie within a certain range of values. When both the number of similar experiments and the number of histogram bars are very large, the histogram becomes, in the limit, a smooth function, which is the probability density. (The histogram and the probability density differ by a constant factor: the total area under the probability-density curve [integral] is, by definition, the total probability of having any value of the mutant frequency and is, of course, equal to 1.) The density function contains information about the most relevant statistical parameters (average values and standard deviations) which can be compared with experiment (see "Experiments on evolution and observable parameters" below). In particular, the characteristic width of the probability density peak indicates the error within which the mutant frequency can be predicted.

The stochastic evolution equation (equations 1 and 2) (Fig. 2a) expresses the rate of change in the probability density with time in terms of its form at the present moment. Using such an equation and knowing the initial probability density, one can predict its form, in principle, at any time in the future, similarly to how one would predict the mutant frequency itself for a deterministic process. The difference between the two cases is that the time-dependent variable is now a function rather than a number. We derive the evolution equation directly for the population model introduced in the previous subsection, in the beginning of mathematical part of our work (see "Mathematical results and derivations" below). The rest of the mathematical part is devoted to solving the equation for different important cases. Here we only show how the equation looks when the probability density is localized in a small region near some value of the mutant frequency and comment on its meaning from a more qualitative perspective.

The right hand-side of the equation shown in Fig. 2a is a

sum of three terms, which together describe how the shape of the probability density function, ρ, changes over a short time interval, *dt*. The first term describes random drift, the second describes selection, and the third describes mutation. To clarify the roles of the three terms in describing evolution, we consider each of them separately, by setting the other two terms equal to 0 (Fig. 2b to d). As a convenient example, we examine a probability density localized in a small region near some value of the mutant frequency (f_{max}). In this example, the second term, by itself, means that the probability density increases with time on the left side of the peak and decreases on the right side of the peak. As a result, the probability density peak, whose shape stays constant, shifts to lower mutant frequencies, as it should in the presence of selection (Fig. 2c). The third term in the equation, by itself, causes a shift of the peak as well, but the direction of the shift is toward 50% composition, which is the expected effect of mutation when the forward and reverse mutation rates are, as assumed, equal (Fig. 2d). The effect of the first term in the equation is of a different kind. Due to this term, the probability density decreases in the interval between the inflection points A and B (Fig. 2b) and increases everywhere outside of the interval. As a result, the probability density spreads outward. This is random drift: the error within which one can predict the value of mutant frequency increases with time. A more general form of the stochastic equation when the probability density, ρ(*f*), is spread over a broad interval of *f*, is given in equations 1 and 2.

In the equation in Fig. 2a, a physicist will recognize a particular case of the Fokker-Planck equation and a mathematician will recognize a case of the forward Kolmogorov equation (41). It was introduced into the field of population genetics by Wright (81) and then intensively used to study evolution in the presence of different factors (31–33, 37). As it turns out, the equation is much more general than the virus model we used for its derivation in the mathematical section of this review. It describes a broad range of population models, from a bacterial culture to a randomly mating population without allelic dominance (35). Originally, the approach of the Fokker-Planck equation was introduced into population genetics from a phenomenological perspective, based on analogy to gas kinetics (18). Later, the validity of this approach was confirmed for different population models (52, 75). Examples of essential factors which are not included in the equation but which may or may not be important, depending on the experimental system, are epistasis (biological interaction) and linkage between multiple loci, time variation of the selection coefficient and the population size, and allelic dominance in a diploid population (33).

A formal analogy for the system described by the evolution equation is a gas consisting of particles mixed with air and confined between two parallel walls (Fig. 3a). A value of the mutant frequency is analogous to a location between the walls, and the probability density is now the local gas density. The first term (Fig. 2a) describes the diffusion of the gas particles in the air, and the second and third terms combined describe the effect of directed force (an electric field, for example) acting on the gas particles in the presence of friction of the gas against the air. Another useful analogy is gel electrophoresis. The electrical force acting on polymer molecules and the friction against the gel matrix together create directed motion, which segregates the molecules into bands. Molecular diffusion leads to increasing bandwidths. Although the physics of the gel or gas system has nothing to do with viruses or evolution, the formal mathematical analogy between the two systems, as we shall see below, turns out to be very useful.

Boundary conditions: properties of almost monomorphic populations. In the real world, the mutant frequency cannot be less than 0 or greater than 1, yet the master equation has no such restriction. Thus, the stochastic equation in Fig. 2a (and equations 1 and 2) is incomplete without describing what happens near ends of the allowed interval for the mutant frequencies, 0 and 1. The analysis shown in Fig. 2 is for the case where there is a large number of minority allele copies (that is, *f* is not near 0 or 1) and treats the mutant frequency (*f*) as a continuous variable. In many important cases, one also needs to describe the evolution of a population with only a few copies of the minority variant. The boundary conditions where *f* is near 0 and 1 have to be derived independently from the virus population model described in Subsection A. The derivation given in the mathematical section of this review shows that the conditions differ depending on the interval of population size, as follows.

The boundary conditions can be conveniently expressed in terms of the probability density flux (*q*), which is exactly analogous to the flux of gas particles through unit area per unit time (Fig. 3). In very large virus populations (Fig. 3b), the boundary conditions state that the flux must vanish at the "walls" corresponding to two monomorphic states, i.e., 100% mutant or 100% wild type (equation 3). In small populations (Fig. 3c), the flux is not zero (equations 5 and 6). This is because the probability of finding the virus population in a completely monomorphic state is finite and can increase or decrease in time. In the gas analogy, in the first case (Fig. 3b) gas molecules bounce off the hot walls and in the second case (Fig. 3c) the walls are cold and gas forms a condensate which can decrease or increase with time. Figuratively speaking, the probability density, just like the gas condensing in or evaporating from the liquid on a wall, can "condense" in or "evaporate" from a monomorphic state.

The real, biological interpretation of the different sets of boundary conditions is as follows. In very large virus populations (which, as we shall see, roughly correspond to almost deterministic evolution), a purely monomorphic state is unlikely: mutations destroy it very quickly. In a small population, mutations are rare and the monomorphic state can occur with a finite probability. This argument also shows that mutations affect virus evolution in a different way depending on the number of infected cells. In a large population, mutations may be important even in a very polymorphic state (e.g., if selection is small). In small populations, the role of mutations is to create a copy of the new allele in an otherwise monomorphic population; once a copy is created, mutations can be neglected until the population becomes monomorphic again. Typically, as we discuss below in the section on steady state, a new allele is lost due to random drift and repeated introduction of mutations will be needed to restore diversity.

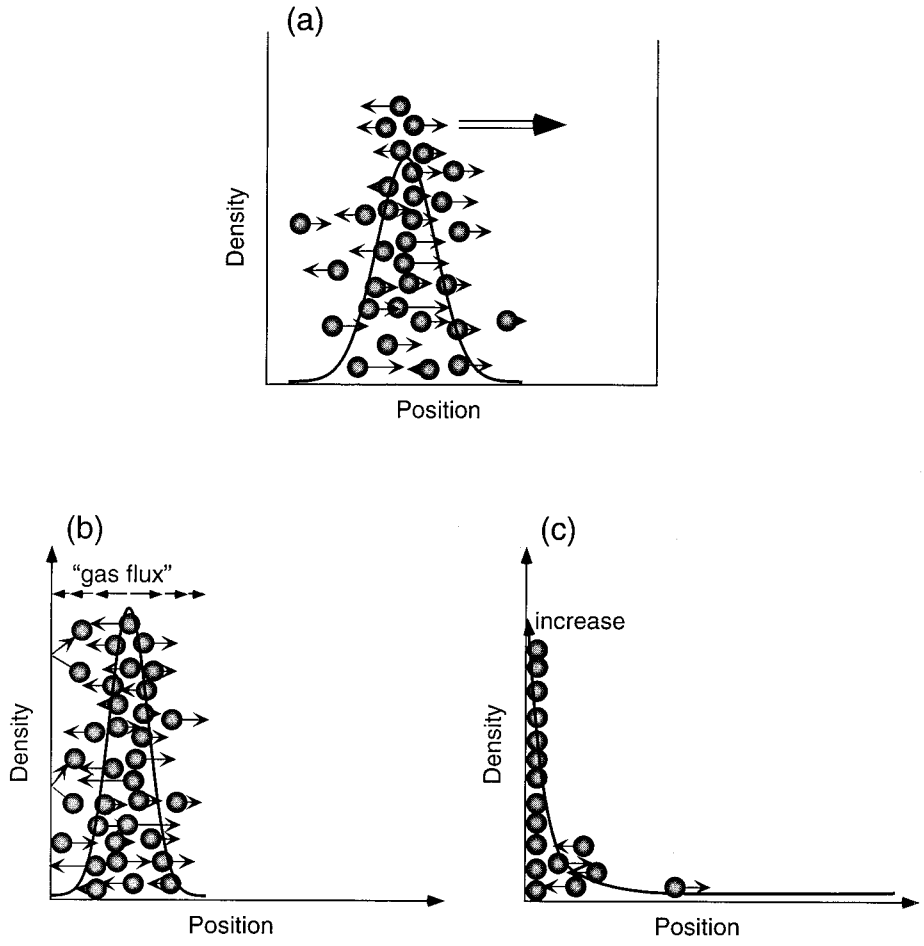

FIG. 3. Stochastic evolution equation (equations 1 and 2) and its boundary conditions viewed through the formal analogy between the probability density and the local gas density. The walls at $f = 0,1$ correspond to the two monomorphic states. (a) Gas particles subject to diffusion and a directed force when far from the walls. (b) Boundary conditions at large population sizes: gas particles bounce off the walls; the total flux at a wall is 0 (equation 2). (c) Small population sizes: gas particles can condense on or evaporate from a wall; the total flux at a wall does not need to be 0 (equations 5 and 6).

Experiments on Evolution and Observable Parameters

In this section, we describe a few gedanken experiments on genetic evolution important for virological applications and introduce quantitative parameters suitable for experimental comparison.

To make use of the evolution equation with boundary conditions (see "Description of the model and the evolution equation" above), one needs to know the state of the system or its statistics at the initial moment of time. The initial condition depends on a particular experimental or natural setup. Virological experiments, relevant for both in vivo and in vitro situations, are as follows.

(i) Accumulation of deleterious mutants (initial condition: a pure wild-type population, i.e., $f = 0$).

(ii) Reversion of a deleterious mutation (initial condition: a pure mutant population, i.e., $f = 1$).

(iii) Growth competition (initial composition: a 50%-50% population [$f = 0.5$] or any other strongly polymorphic mixture).

(iv) Gene fixation (this experiment, which has received a lot of attention in population biology [19, 24, 34, 38, 80] and which is very useful for understanding other stochastic experiments, is defined only in small populations in which the total mutation rate per population, μN, is much less than 1; suppose that a single advantageous allele is introduced into an otherwise monomorphic population [$f = 1/N$]—the allele will have one of two fates: either it will be lost due to random drift [Fig. 1a] or it will spread to the entire population, i.e., become "fixed"; the questions are: what is the fixation probability, and, if the allele is fixed and does not become extinct, how much time will it take, counting from the moment it appeared? One can also ask a more general question: what is the probability of having a new allele to grow into a subpopulation of a given size before it becomes extinct?).

(v) Steady state. Whatever the initial condition, after a sufficient time, the system passes to the stochastic steady state, in which the probability density no longer depends on time; we consider this relatively simple case separately.

(vi) Genetic divergence. One splits a steady-state population into two isolated parts. Initially, both populations have a random but identical genetic composition, from which they independently diverge. As time goes on, their respective random

compositions correlate less and less. The question is, what is the characteristic time at which the loss of correlation occurs?

(vii) Genetic turnover? This experiment studies the average timescale associated with random fluctuations of the mutant frequency in the steady state.

The probability density (ρ) of the mutant frequency predicted by the stochastic equation is the main observable parameter. Unfortunately, to measure it directly, one would have to generate a histogram of mutant frequencies for a very large ensemble of populations. More amenable for experimental testing are the average (expectation) values (equation 36) and the standard deviations or variances (equation 37) of different stochastic parameters, which require a smaller number of populations to measure. Below we introduce some useful parameters whose statistics can be measured in the different experiments we outlined above. At the same time, their predicted statistics can be expressed via the probability density, as shown in the mathematical section of this review. In what follows, we assume that each parameter, for each given population, is measured with a high precision from a sufficiently large sample of sequences. The sampling effects will be discussed separately below.

The first parameter is the mutant frequency itself (f), which is self-explanatory. Its value can be compared directly with the experimental value, provided that the wild-type (best-fit) nucleotide is known.

The second is the intrapopulation genetic distance (T), defined as the proportion of sequence pairs (randomly sampled from the virus population) which differ at the base of interest. Although there are other ways to measure intrapopulation variability, we will use this definition, known in population biology as Nei's nucleotide diversity. It is equivalent to the standard definition of the genetic distance in virology as the average number of pairwise differences among randomly selected genomes, except that it applies to a single base rather than to a long genomic segment. By definition, T is calculated as $2f(1 - f)$, and varies between 0 (at $f = 0$ or 1) and 0.5 (at $f = 0.5$). The genetic distance is usually a more convenient measure of population diversity than the mutant frequency itself since it does not require knowledge of the wild type sequence.

The third is the interpopulation genetic distance (T_{12}), which is defined in the same way as the intrapopulation genetic distance, except that the two sequences of each pair are sampled from two different populations (equation 40). The interpopulation distance is 0 when the two virus populations consist uniformly of the same genetic variant and 1 (100%) when the two virus populations are composed entirely of opposite genetic variants. The interpopulation distance, as one can show, cannot be smaller than the average of the two intrapopulation distances. Therefore, it is sometimes more convenient to consider instead the relative genetic distance between two populations (D), defined as the difference between the interpopulation distance and the average of the two intrapopulation distances [$T_{12} - (T_1 + T_2)/2$]. This parameter (equation 41) varies between 0 (two populations have an identical genetic composition) and 1 (one population is pure mutant, another is pure wild type). There are alternative definitions of the relative distance (54). We find this definition more clear intuitively; also, its statistical moments (average, variance) are relatively easy to calculate.

All the previous parameters can be measured at one time point, both for dynamic experiments (the first three experiments in the beginning) and in the steady state. Since all of them are, in general, stochastic, an average and standard deviation has to be calculated for each. The next parameter is more complex: it requires measurement at two different times. We define it on average and for a steady state population only.

The fourth parameter, the time correlation function of mutant frequency [$K(t)$], describes how quickly the system "forgets" the preceding random fluctuation of the mutant frequency (equation 45). The time correlation function usually has a maximum when the time difference is 0 and vanishes at large time differences. The characteristic time at which it decays by 50% (or, say, by a factor of $e = 2.78\ldots$) from its maximum gives the timescale of random fluctuations. The form of this decay (e.g., exponential or negative power) may be a good fingerprint of a virus population model or, within a given model, of a particular population size.

In the mathematical section of this review, we calculate these parameters for different gedanken experiments and different intervals of population size. In this section of the review, we discuss these results qualitatively and illustrate them, when possible, with Monte Carlo simulations.

Steady State

In this section, we discuss properties of the steady-state, stochastic population in different intervals of the population size.

Neutral case: $s \ll \mu$. Selection is of little significance when the selection coefficient is much less than the mutation rate. This case is probably of little practical significance for RNA viruses, with their tightly organized genomes. However, the transition between stochastic and deterministic behavior is easier to analyze when the selection factor can be neglected. Hence we start our discussion here.

The main fact of stochastic theory is that fluctuations of mutant frequency between statistically identical populations are large if populations are small (stochastic behavior) and small if populations are large (nearly deterministic behavior). In the language of the probability density (equation 52), the density is spread over a broad interval of f in small populations and is a narrow peak at very large population sizes. Transition between the two limits is controlled mostly by a single parameter μN, the product of the population size and the mutation rate. The composite parameter μN, which features extensively in population genetics (usually as $\Theta = 2N\mu$), gives the total mutation rate for the entire population. For most RNA viruses, μN equals 1 when the number of infected cells is on the order of 10^5 (i.e., less than the number in a small culture dish).

As the mutation rate per population increases, the probability density gradually changes its shape, as illustrated in Fig. 4 (80). This results from competition between random drift, which drives the system to one of uniform states, and mutations, which diversify the system. At values of μN much smaller than 1 (an interval we accordingly call the drift regime in Table 1), random drift wins and the usual population is only weakly polymorphic. The probability density is, accordingly, U shaped, with a minimum at 50% composition. At the smallest values of μN (the condition is given in equation 5), the system is most likely to be in either of the purely monomorphic states, without a single opposite allele present (see "Description of the model

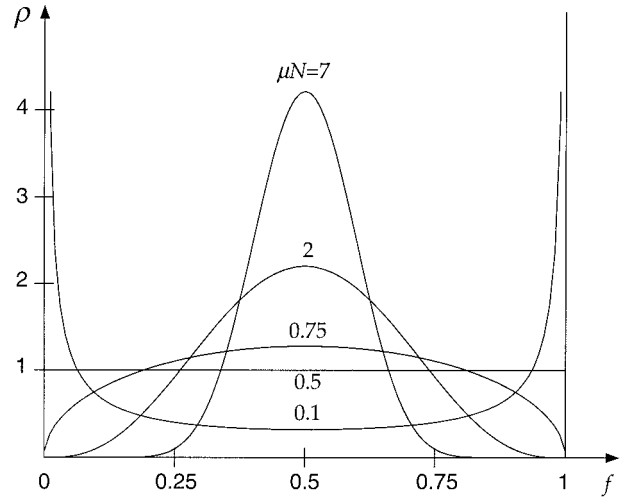

FIG. 4. The steady-state probability density in the neutral case. The curves show $\rho_{ss}(f)$ when $s \ll \mu$ at different population numbers, N. Numbers on the curves show the corresponding values of μN.

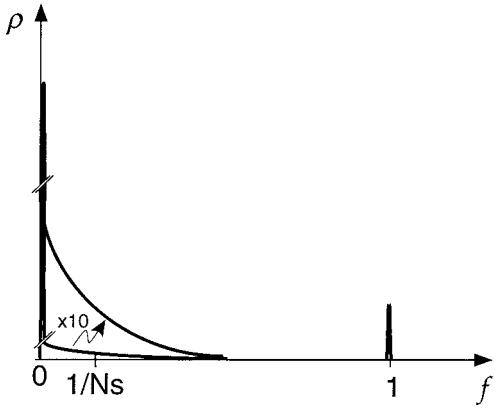

FIG. 5. Schematic plot of the steady-state probability density in the selection-drift regime. The curve shows $\rho_{ss}(f)$ for the case when $1/s \ll N \ll 1/\mu$. Note the very narrow peaks at $\rho = 0$ and 1, together with the tail extending from $\rho = 0$.

and the evolution equations" above, where the the boundary conditions are described). The total probability of any polymorphic state will be much less than 1 and on the order of μN. This estimate gives the frequency of segregating sites in a genome segment.

Let us move toward larger populations. As we increase the parameter μN, the U shape of the probability density flattens out (Fig. 4). The minimum at 50% composition becomes a maximum when μN is equal to 1/2. The probability density shrinks and becomes narrow as the population increases and μN becomes much larger than 1. This means that the mutant frequency is very close to the deterministic value of 1/2, owing to the balance between forward and reverse mutations. In Table 1, this limit of population sizes is denoted the mutation regime.

Case with selection: $\mu \ll s \ll 1$. The situation when the selection coefficient is less than 1 but still much larger than the mutation rate is more relevant for RNA viruses and more interesting theoretically. As in the neutral limit, the larger the population size the smaller the fluctuations.

The selection factor can be neglected only if a population is very small, much smaller than the inverse selection coefficient ($Ns \ll 1$), a case that has the same properties as the above-described drift regime. At larger population sizes, selection is crucial and causes the probability density (equations 48 or 49 to 51) to be asymmetric in favor of a predominantly wild-type population.

In the limit of very large populations, when μN is much larger than 1 (termed the selection regime in Table 1), the probability density is narrow and localized near its deterministic value (equation 57). This value is given by the ratio of the mutation to the selection rate (μ/s), which we assumed to be small. At this value, mutations and selection against emerging mutants reach balance.

A result not sufficiently emphasized in the population biology literature is the existence of a wide interval in population size between the inverse mutation rate and the selection coefficient, which we term the selection-drift regime, in which all three factors of evolution are critical. Specifically, mutations produce diversity, selection restricts mutants to a low level, and random drift causes strong fluctuations between populations. The structure of the probability density in this regime is shown schematically in Fig. 5. It consists of three components. The large peak (delta function) situated at exactly zero mutant frequency means that a population is, most probably, purely wild type. The weak continuous exponential tail which decays at mutant frequencies on the order of $1/Ns \ll 1$ (80) means that the chance of a population being polymorphic is low and that if a population happens to be polymorphic, the proportion of mutants is small and quite random. A small peak at $f = 1$ becomes important only close to the lower border of the interval, when N is on the order of $1/s$. The probability of finding any mutants (which is given by the total area under this curve) is low and proportional to μN (equations 49 to 51).

The selection-drift regime has rather interesting, even controversial properties. On the one hand, the shape of the prob-

TABLE 1. Classification of regimes of genetic evolution

Regime	Neutral limit ($s \leq \mu$)				In the presence of selection ($s \gg \mu$)			
	Population size	Behavior	Factors in steady state	Factors in diverse state	Population size	Behavior in steady state	Factors in steady state	Factors in diverse state
Drift	$N \ll 1/\mu$	Stochastic	Drift, mutations	Drift	$N \ll 1/s$	Stochastic	Drift, mutations	Drift
Selection-drift					$1/s \ll N \ll 1/\mu$	Stochastic	Drift, mutations, selection	Selection
Selection					$N \ll 1/\mu$	Deterministic	Mutations, selection	Selection
Mutation	$N \gg 1/\mu$	Deterministic	Mutations	Mutations				

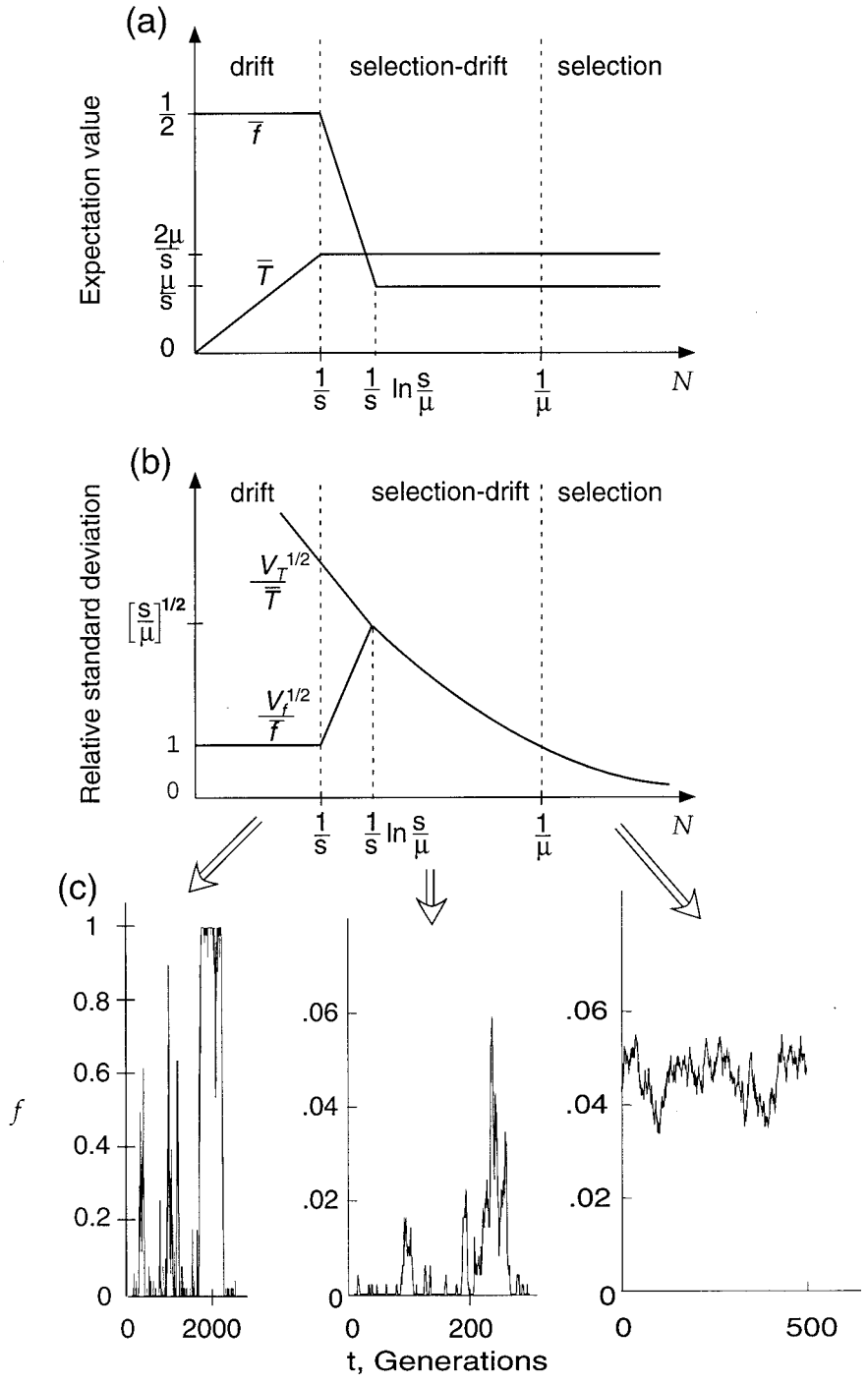

FIG. 6. Dependence of the observable parameters at steady state on the population number. N varies over the three main intervals. (a) Average mutation frequency, $\bar{f}$ and genetic distance, $\bar{T}$. (b) Relative standard deviations of the same two parameters. (c) Fragments of representative Monte Carlo simulations in the respective intervals of N (see Fig. 10 to 12 for details).

ability density suggests a very stochastic behavior. On the other hand, the average mutant frequency and the average genetic distance happen to coincide, over most of the regime, with their deterministic values, as if the population were much larger. Figure 6 shows the average values and the relative standard deviations for both parameters at all the population sizes. As expected, in the selection drift regime the relative standard deviations for both the mutant frequency and the genetic distance are much larger than unity (Fig. 6b). At the same time, the average values (in equation 59) are the same as in the selection regime (Fig. 6a). Notably, the fluctuations of the parameters are much stronger than could be expected from the Poisson statistics. This is a result of clonal amplification: if a single mutant appears in otherwise wild-type population, it

(a) $\rho(f,t) = \delta(f - f_d(t))$

$$\frac{df}{dt} = -sf_d(1-f_d) - \mu(2f_d - 1)$$

(b)

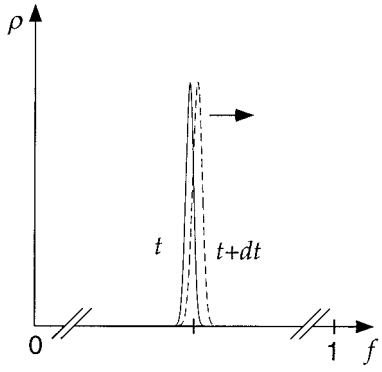

FIG. 7. Probability density of the mutant frequency in the deterministic limit. $\rho(f)$ is represented by the mathematical expression for $\mu N \gg 1$ (a) and a schematic plot (b).

grows into a clone. In the sections on stochastic dynamics (see below), we will further clarify the structure of the steady state by presenting a Monte Carlo simulation of a stochastic dynamic evolution in a single population. Examples of the results of such simulations for each regime are shown in Fig. 6c.

Deterministic Dynamics and Its Boundaries

As we have shown above (see "Experiments on evolution and observable parameters"), the steady-state mutant frequency approaches its deterministic value when μN is much larger than 1. The purpose of this section, small but with a large mathematical counterpart, is to gain insight into the transition between stochasticity and determinism in the more complex case, in which parameters of the system depend on time.

Deterministic dynamics. Deterministic and stochastic theories operate with different dynamic variables. The former considers the time dependence of the frequency of mutants, and the latter uses a more complex object, the time-dependent probability density of the mutant frequency. It is important to ensure that the two approaches converge to the same result in the limit of infinite population, when they are expected to describe deterministic evolution, albeit in a different way. For this purpose, in the mathematical section of this review we solve the dynamic stochastic equation (equation 1) for the case of large populations. The resulting probability density, as expected, is a very narrow peak located at the time-dependent mutant frequency (Fig. 7b), which satisfies the deterministic equation of evolution (equations 60 and 61).

The first term in the right-hand side of the deterministic equation (Fig. 7a) (equation 61) describes selection for the wild type, causing depletion of mutants. When one of two subpopulations (f or $1 - f$) is very small, the first term becomes small, since if there is no diversity, there is no selection. The second term, describing mutations, does not vanish in a uniform population. Instead, the term vanishes at 50% composition when the effects of forward and reverse mutations cancel each other. Mutations drive the system toward 50% composition. The same evolution equation can be obtained directly from the deterministic first principles (equations 63 and 64).

The deterministic equation in Fig. 7a allows one to predict the genetic composition as a function of time for any initial condition set in an experiment (equation 62). Corresponding plots for the three cases matching the conditions of the accumulation, growth competition, and reversion experiments described above (see "Experiments on evolution and observable parameters") are shown in Fig. 8. In all cases, after a characteristic time proportional to the inverse selection coefficient ($1/s$), the population approaches a steady state in which the mutant frequency saturates at a small value, the mutation rate over the selection coefficient (μ/s) (see "Steady state" above). Reversion is somewhat delayed compared to that in the two other experiments since the system first has to diversify slowly due to mutations and then still has to cross the entire interval of the mutant frequencies. Note that in both the accumulation and reversion experiments, the initial slope of the time dependence of the mutant frequency is shallow and is determined by the mutation rate (Fig. 8). Selection becomes important and causes the plots to curve after a growing subpopulation becomes sufficiently large.

Boundaries of deterministic approximation. Random drift, always present even in very large populations, causes the frequency of mutants to fluctuate around its deterministic value. As the population size decreases, the magnitude of fluctuations becomes comparable to the average frequency of the minority

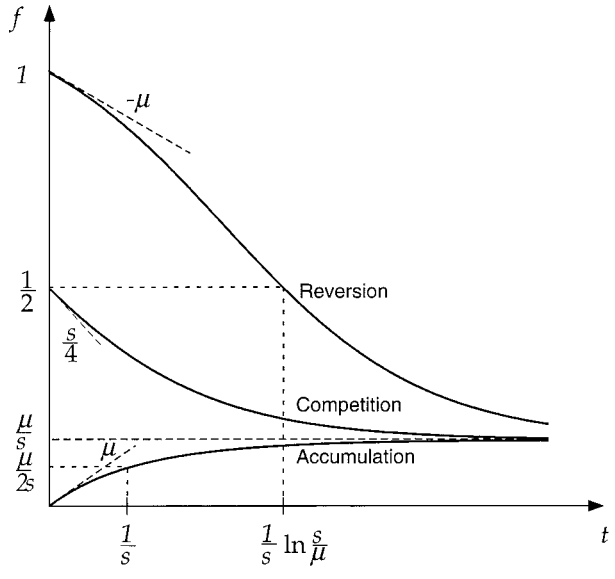

FIG. 8. Schematic dependence of the mutant frequency, f, on time in the deterministic limit. The three curves correspond to three different initial values of $f(0)$: accumulation of mutations [$f(0) = 0$], growth competition [$f(0) = 1/2$], and reversion of a mutation [$f(0) = 1$]. The value of the ratio μ/s used in the figure is unrealistically high for viruses and is used for clarity of plot only. Dashed lines show initial slopes.

allele (either mutant or wild type), and the deterministic description breaks down. The corresponding condition on the population size varies significantly depending on the initial conditions of the experiment (equation 65). When the population starts from a monomorphic state (reversion or accumulation), the deterministic criterion is met when μN is much larger than unity. A population that is strongly diverse to start with, as in the growth competition experiment, is already deterministic at a much smaller population size in the selection-drift regime. (The criterion for diversity is that the mutant frequency must be higher than its characteristic "tail" at steady state [Fig. 5]). The reason for this difference is that a small polymorphism is influenced by rare and random mutation events while a strongly polymorphic population is controlled by selection alone.

Stochastic Dynamics: the Drift Regime

At the smallest population sizes, smaller than the inverse selection coefficient, as we found out when considering the steady state, selection can be neglected altogether. In this section, we consider the nonequilibrium dynamics in this regime. The problems of interest are those listed above (see "Experiments on evolution and observable parameters"): the decay of a strongly polymorphic state, gene fixation, transition from a monomorphic to the steady state, divergence of populations which have been separated, and the rate of genetic turnover in the steady state.

Decay of the polymorphic state and gene fixation. We start our discussion from the population that is initially polymorphic, somewhere in the middle between 0 and 100%. As already discussed (see "Description of the model and the evolution equation"), mutations are not important in a polymorphic population, since they occur in the population with a frequency, μN, much less than 1 per generation. Therefore, random drift remains the only factor causing variation of the mutant frequency in time. As time passes, the mutant frequency drifts until the population accidentally ends up in either monomorphic state (cf. Fig. 1a). A representative random process is illustrated by computer simulation in Fig. 9b. The average time (the number of generations) it takes for a population to become monomorphic (i.e., for either variant to be fixed) is on the order of the population size (equations 81 and 82) (32, 80). The fixation time is quite random: its representative fluctuations are on the order of its average value. The same process can be understood in another way, from the time evolution of probability density. Figure 9a shows how the probability density, initially a narrow peak located, e.g., at 50% composition, gradually spreads out to the entire interval and then decays.

The fact that, in a time not exceeding a few multiples of the population size, the population becomes uniform has general phylogenetic consequences. Let us divide arbitrarily a population into two groups of equal size and mark each group, say, by a different color. Then we divide each group (color) into two subgroups and mark them by two different shades. Then we divide each shade into two hues, and so on. If we continue the process of subdivision long enough, all individuals in the population will eventually have different tags. Consider now a group consisting of two subgroups. According to the above

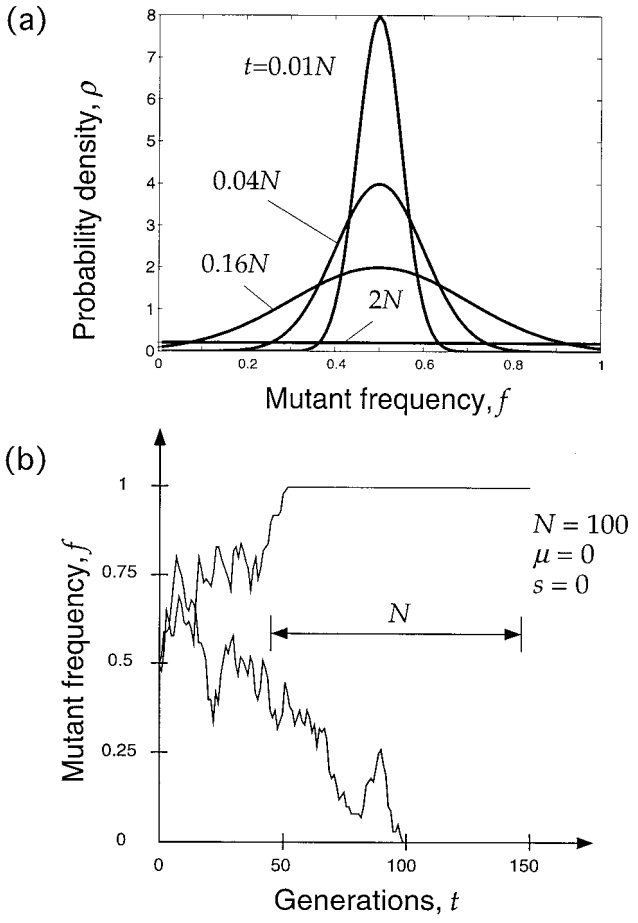

FIG. 9. Decay of polymorphism in the drift regime. A growth competition experiment for the initial condition $f_0 = 0.5$ and $Ns \ll 1$ is shown. (a) Change in the probability density in time (equations 81 and 82). (b) The two stochastic dependences $f(t)$ were obtained by random runs of a Monte Carlo simulation program written for the virus population model described in the text. Parameters are shown in the figure.

result, in a time not exceeding a few multiples of the group size, one of the two subgroups vanishes. Likewise, the surviving subgroup contains two smaller subgroups, one of which also becomes extinct in a time not exceeding a few multiples of the subgroup size, and so on. Therefore, in a time on the order of the total population size, the entire population will have the same tag, i.e., will comprise descendants from a single virus or organism. In other words, any two organisms in a population in the drift regime have a common ancestor at a past number of generations on the order of the population size. Phylogenetic methods of analyzing branching processes confirm this result, which is the basis of the coalescent method of estimating population size (39, 40, 65).

Related to the decay of polymorphism described above is gene fixation. Suppose that a single new allele is introduced into a monomorphic population at an initial moment. Eventually, after a number of replication steps, the allele will either disappear due to random drift (which is the most likely outcome) or spread to the entire population, i.e., become fixed. The questions are as follows. (i) What is the probability that

the allele will get fixed? (ii) Given that the allele is lucky enough to become fixed, what is the average fixation time? As we show in the mathematical section of this review (equation 84 with $f = 1$), the fixation probability is the inverse of the population size ($1/N$) (34) and the fixation time is on the same order as the polymorphism decay time, i.e., on the order of the population size.

One can also ask more general questions. What is the probability that a single mutant genome will ever grow into a subpopulation with a given size? What is the average time spent on this growth? The results are analogous to that for full fixation, except that the subpopulation size substitutes for the total population size (equations 84). As we show in the beginning of the sections on stochastic dynamics in the mathematical section of this review, this result allows us to interpret, at a semiquantitative level, all the important results on stochastic dynamics.

Transition from a monomorphic to a steady state. We also consider here the accumulation of mutations starting from a purely monomorphic state, e.g., wild type (which one of the two does not matter, since selection is negligible). Eventually, mutants will be generated, one of them will become fixed (as described), and the system will switch to pure mutant. Then wild-type alleles will be generated, etc., and, in the long run, the population will be, statistically speaking, in dynamic steady state in which it switches back and forth between two monomorphic states. The system will gradually "forget" its initial state, so that the probabilities of the two monomorphic states will be equal and will be close to 1/2.

In the probability density language, this process can be described as shown in Fig. 10a. The initial peak of the probability density is very narrow and is localized at the zero mutant frequency. As time goes on, a tail of the probability density spreads into the interval between 0 and 100% mutants (equations 85 and 86) and a new peak at 100% mutants appears, reflecting a chance of early fixation of a mutant genome. The first peak decays and the second peak grows, until they become equal in the steady state (Fig. 4) (equation 87). In the gas system analogy (see "Experiments on evolution and observable parameters" above), all water is initially condensed on the left wall and then evaporates. The vapors diffuse into the container and condense again on the right wall (analogous to what happens in a freezer over time). The system reaches equilibrium when the amount of condensate on both walls is the same and there remains some gas in between.

In addition to the language of probability density, it is useful to visualize transition to the steady state directly, as a typical random process. If the probability density is analogous to the density of gas, the random dependence of the mutant frequency on time corresponds to the random trajectory of a separate gas particle. A representative Monte Carlo simulation of the equilibration process, together with the relevant timescales, is shown in Fig. 10b. The steady-state process looks like a telegraph signal between the two uniform states. The peaks in the mutant and wild-type frequencies correspond to alleles which were generated by mutations and started new subcolonies but failed to become fixed.

Two, widely different timescales appear in both the representative random process and the evolution of the probability density. The typical waiting time for a switch from pure wild

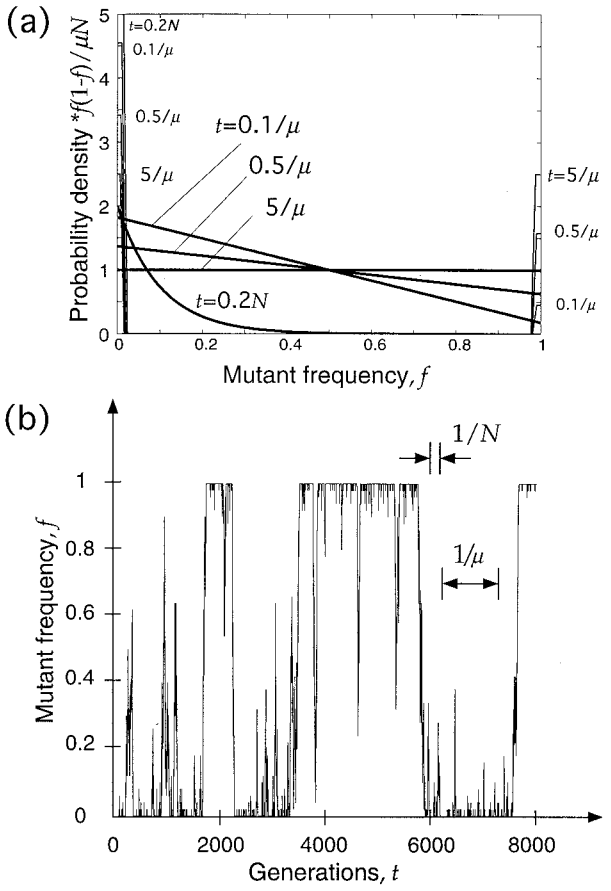

FIG. 10. Time dependence of the mutant frequency in the drift regime on a long timescale. (a) Change in the probability density in time (equations 85 to 87). Sharp peaks at $f = 0$ and 1 correspond to the monomorphic states; their probabilities are shown by the relative peak heights (arbitrary units). (b) One Monte Carlo run is shown for $Ns \ll 1$ and the initial condition $f_0 = 0$. Parameters are shown in the figure.

type to pure mutant or back is within an order of magnitude of the inverse mutation rate $1/\mu$. This corresponds to the time in which the probability density becomes symmetric between the wild type and mutant (Fig. 4) (equations 86 and 87). The actual time spent on a successful switch is much shorter, within an order of magnitude of the population size N. This corresponds to the time in which the tail of probability density is formed between 0 and 100% (equation 85). The two timescales can be derived either rigorously, from the evolution equation (equations 4 to 6), or approximately, from the gene fixation problem (equation 84). Both approaches are used in the mathematical section of this review. They agree with each other and with the simulation in Fig. 10b.

The total probability of a polymorphic state (the frequency of segregating sites in genome) is, at any time, much less than 1 and on the order, roughly, of μN. This agrees with the result we obtained directly for the steady state (see above). Interestingly, this value is reached on a timescale of approximately N generations, i.e., much sooner than the two probabilities of monomorphic states equilibrate.

Divergence of populations which have been separated and the time correlation function. The longer timescale, $1/\mu$, also

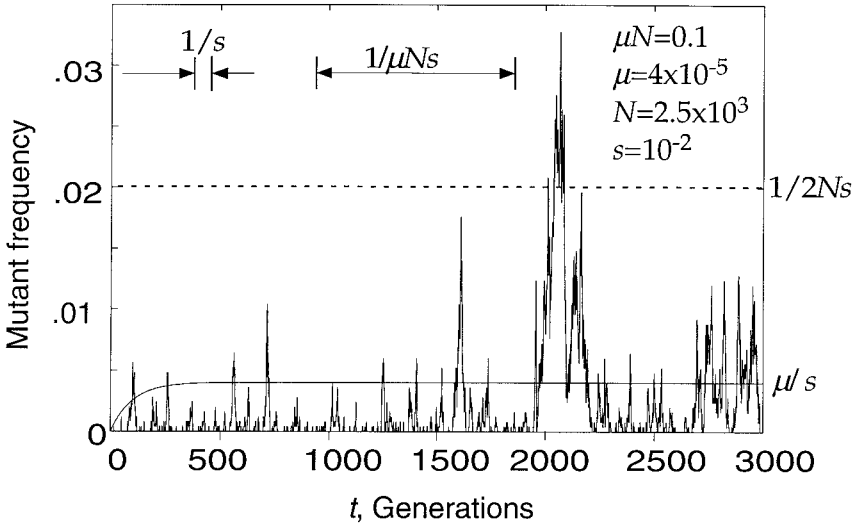

FIG. 11. Simulated accumulation of mutants in the selection-drift regime. One random Monte Carlo run is shown for $1/s \ll N \ll 1/\mu$ and the initial condition: $f_0 = 0$. The double-pointed arrows and dashed line show predicted scales in time and in the mutant frequency. The solid smooth line shows the deterministic dependence for comparison. Parameters are shown in the figure.

appears in the time correlation function of mutant frequency, which characterizes the timescale of random fluctuation in the steady state and the divergence of populations which have been separated (see "Experiments on evolution and observable parameters" above). The value of the relative genetic distance, D, gradually changes from 0 to a constant value corresponding to statistically independent populations (equation 90). (Note that some other measures of interpopulation genetic distance used in population biology do not have an upper limit [54].) As it turns out, the time of this transition, the half time of the correlation function decay (equation 91), and the time in which the probability density becomes symmetric (above) are on the same order, the inverse mutation rate. Indeed, all three times are determined by the waiting time for a successful gene fixation.

Stochastic Dynamics: the Selection-Drift Regime

Here we consider nonequilibrium experiments in the most interesting interval of population sizes (Table 1). The relative role of selection and stochasticity in population dynamics, as derived from the evolution equation in the mathematical section of this review, depends on the initial genetic composition. The dynamics of growth competition is almost deterministic (see "Deterministic dynamics and its boundaries" above), so that this experiment need not be discussed again. In the accumulation experiment, the overall dynamics is stochastic, except for the average values of the mutant frequency and the intrapopulation distance, which are, remarkably, the same as in the corresponding deterministic conditions.

Accumulation. As in the drift regime (see above), accumulation can be described as a spread of the peak of the probability density initially located at 0 (uniform wild type) into the interval between 0 and 1. However, unlike in the drift regime, the resulting steady state is not symmetric of a large peak (Fig. 5) (equation 48 or 49 to 51). The process of accumulation is reduced to generation of a small tail describing rarely occurring weakly polymorphic states (Fig. 5). As a result, the initial peak at 0 does not decay greatly and the steady state is reached in the same time as in deterministic selection (see "Deterministic dynamics and its boundaries" above) given by the inverse selection coefficient ($1/s$), i.e., faster than all timescales in the drift regime (equations 103 and 104).

The simulated stochastic dependence for this experiment is shown in Fig. 11. The process starts from the generation of a single allele, which tries to grow into a clone. The growth initially occurs under the condition that random drift is more important than selection. The maximum frequency that this clone can reach is determined by the characteristic mutant frequency at equilibrium, $\sim 1/(Ns)$ which corresponds to the clone size, $1/s$ copies (Fig. 5). Above this value, selection becomes the leading force and drift becomes a correction. Further growth of the deleterious clone cannot occur, and it soon becomes extinct. This appears as sparse peaks, the highest of which reach to the length of the "tail" of the probability density, $1/(Ns)$ (Fig. 5) (equation 48 or 50). The half-life of a mutant clone (width of a large peak) is the inverse selection coefficient. Note that the typical time interval between peaks, $1/(\mu Ns)$, is longer than $1/s$. The former time is the waiting time for a new allele that will be lucky to reach the size $1/s$. The latter time is the time that the lucky clone actually spends growing and contracting before it becomes extinct again. The ratio of the two times, μN, gives the probability of finding the population in a polymorphic state (the area under the tail in Fig. 5). As in the drift regime, all these estimates can be obtained from both the evolution equation (equation 101) and the more intuitive gene fixation approach (equation 84). For comparison, simulation of an accumulation experiment in the "selection" regime ($\mu N = 20$) is shown in Fig. 12.

Divergence of separated populations and the time correlation function. The characteristic times of divergence of separated populations (Eq. 105) and the decay time of the correlation function (Eq. 106) are on the order of the inverse selection coefficient, $1/s$. Both experiments show for how long,

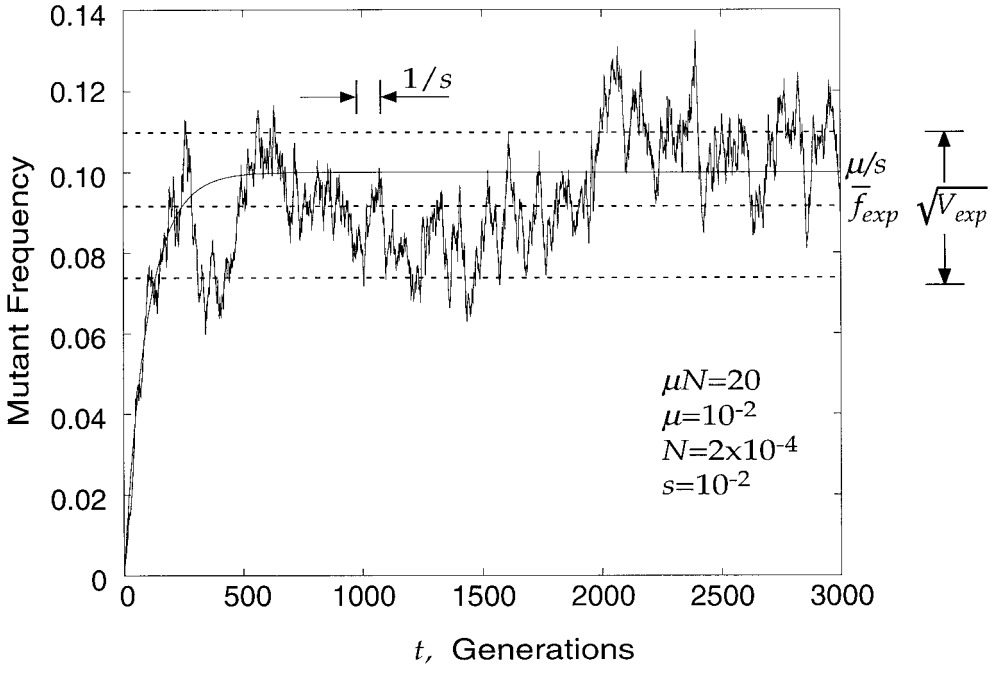

FIG. 12. Simulated accumulation of mutants in the selection regime. Dashed lines show the average and the standard deviation at steady state for $\mu N \gg 1$ calculated using equations 58. Parameters are shown in the figure.

on average, the system "remembers" its previous random fluctuation. The answer: for the half-life of a typical mutant clone, before it becomes extinct. This is because separate clones appear, due to mutation, at independent random times.

Reversion (fixation of an advantageous variant). A reversion experiment, in which the initial population is uniformly mutant, behaves rather differently. Although the same scales for time and the minority allele frequency appear in this case, they have different meaning. As in accumulation, random drift and selection dominate in smaller and larger wild-type colonies, respectively. However, in this case, selection accelerates rather than hinders the growth of a new clone. The probability that a single wild-type allele will manage to grow to a size equal to the inverse selection coefficient, $1/s$, is low, s. However, above this critical size, the rest of its growth will be carried out by selection in a deterministic manner, i.e., with a probability close to 1 and over the deterministic timescale, $1/s$ (see "Deterministic dynamics and its boundaries" above). Hence, the bottleneck of reversion is in reaching the critical size despite random drift; after that, a clone is likely to be fixed in the population. Stochastic dynamics below the critical size is the same as in the accumulation regime (selection is not important). The average waiting time for reversion to start is determined by the fixation probability, s, and by the frequency at which single alleles are generated in a population at each generation, μN, which gives the time $\sim 1/(\mu N s)$, i.e., the same scale as the waiting time for a high peak in accumulation regime (Fig. 11) (equation 107) (51). A few examples of reversion curves are shown in Fig. 13. Evolution of the probability density is shown in Fig. 14, including evolution of the density of polymorphic states (Fig. 14a) (equation 108) and of the two probabilities of monomorphic states (Fig. 14b) (equation 107).

Sampling Effects

In the previous sections, we analyzed random fluctuations of the mutant frequency within an ensemble of populations of infected cells of the same finite size. We have assumed that the value of mutant frequency, genetic distance, etc., for each population was measured accurately by counting the numbers of mutant and wild-type alleles in a sufficiently large sample of genomes. The genome samples used in real experiments are, of course, not infinitely large. Hence, the experimental estimate of any quantity is approximate and sample dependent. The sampling effects may distort the experimental results if the samples are too small. In this section, we calculate how large a sample of genomes we need to achieve a given accuracy of measurements. We focus on the intrapopulation genetic distance (T) defined above (see "Experiments on evolution and observable parameters") for a separate nucleotide. To obtain an experimental estimate of the distance, one isolates a fixed number of sequences from the population, determines the number of nucleotide differences for each pair of sequences, and averages the result over all pairs. (This procedure is specific for our choice of the genetic distance.) The accuracy of the estimate is characterized by the relative error (ε), defined as the standard deviation divided by the average. The result is shown in Fig. 15a (equation 116). This formula is quite general and can be applied to any regime or particular experiment on genetic evolution. For instance, for the maximum possible intrapopulation distance, $T = 0.5$ (in absolute units), which corresponds to the half-and-half variant composition at the base of interest, 25% accuracy is approximately reached at a sample size of 6 genomes and 10% accuracy is predicted for a sample of 14 genomes. As the polymorphism decreases, the sample size required increases quickly. To reach, e.g., a 20% accuracy

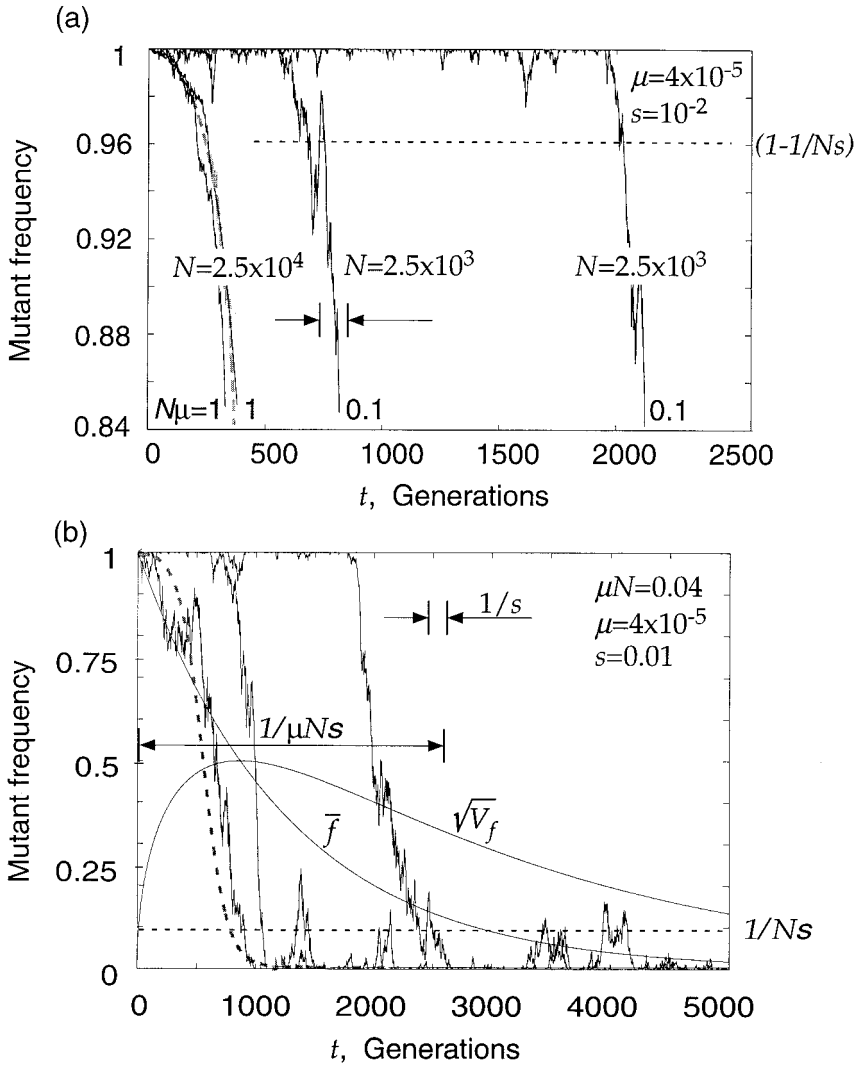

FIG. 13. Simulated reversion (fixation of advantageous variant) in the selection-drift regime, $1/s \ll N \ll 1/\mu$. The reversion curve in the deterministic limit, $N = \infty$, is shown by the dashed curves for comparison. Parameters are shown in the figures. (a) Beginning of the reversion curves. Two random Monte Carlo runs are shown for each of two population sizes. (b) Full reversion curve at a smaller population size. Three random runs are shown. Solid lines show the average and the standard deviation of the mutant frequency calculated using equation 107.

of measurement at the genetic distance $T = 0.095$ (0.95 and 0.05 composition at the base), one needs to sample ~500 genomes (of which 25 ± 5 genomes will be mutant). Hence, to study rare genetic variants, it is undesirable to simply count sequences: one needs to employ alternative methods of quantitation like selective PCR. Of course, as is done often, the genetic distance can be averaged over a large number of bases; this saves the sample size. Such a simple solution will not work, however, if one does not know whether the bases used for averaging evolve under similar conditions or if one is interested in a specific base.

Experimental design requires making an educated prediction of the appropriate sample size and measurement methods, and one therefore needs to anticipate the intrapopulation genetic distance, at least to within an order of magnitude. At the same time, the actual value of the distance fluctuates between populations and is not known before the measurement is made. Therefore, one has to use some sort of theoretically predicted typical distance. Making such a prediction is not trivial. The expectation value is not a good choice, since, deep into the stochastic regime, a population is most probably found in a monomorphic state at any given allele. The sample size has to be optimized with respect to polymorphic states. These states have a low probability: if a population is completely uniform at a site, any size sample will be monomorphic as well. We propose to use the representative average distance (T_{rep}), which differs from the standard average distance in that it is averaged over polymorphic states only. (Experimentally, this can be accomplished by examining many sites and focusing attention only on the few that are polymorphic.) Quantitative differences between the two averages can be rather large. The expressions for the representative average distance in the steady-state population, for all three intervals of the population size, are shown in Fig. 15b (equation 118). One can see

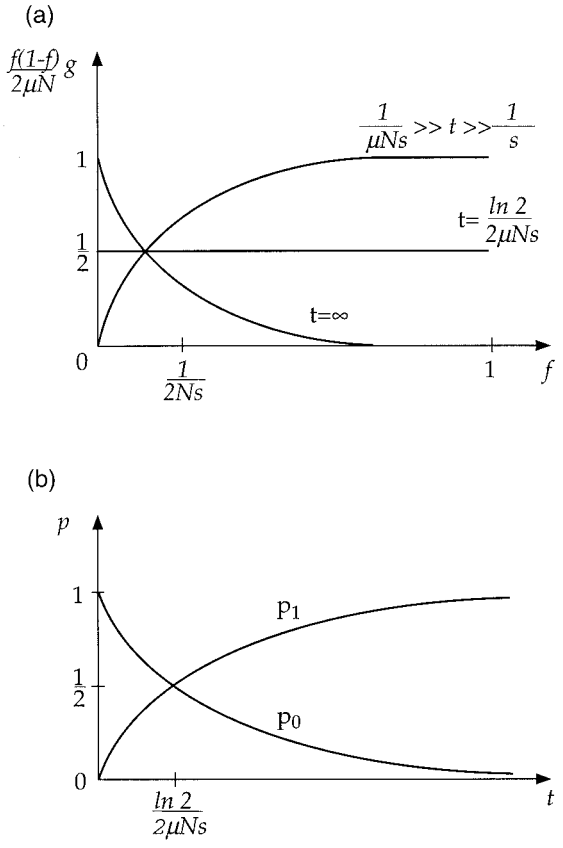

FIG. 14. Reversion (fixation of advantageous variant) in the selection-drift regime, $1/s \ll N \ll 1/\mu$. (a) Evolution of the probability density (equation 108). (b) Probabilities of monomorphic states, p_0 and p_1 (equation 107).

that the smallest distance and therefore the largest samples required correspond to the deterministic limit and the smallest samples correspond to the drift regime. This advantage of stochastic regimes is, however, canceled by a large number of different populations (or, at least, similar sites) needed for a representative assessment of polymorphism (roughly $1/\mu N$ populations or sites). In the steady state, assaying many populations or sites can be traded for sampling the same population or site at many time points spaced farther than the genetic turnover time (see the discussions of the time correlation function in the previous two sections).

Experimental Applications

The theoretical considerations presented here have useful and important implications for understanding the evolution not only of viruses but also of organisms generally. Their practical application, however, requires the use of appropriate experiments, designed both to test the validity of the theory and to then apply it to specific situations. In the following subsections, we present three examples of such applications. Other important experimental issues which are outside of the scope of our basic analytic review (phylogenetic studies, multilocus effects, etc.) will be briefly discussed in the next section.

Virological studies in vitro. Viruses replicating in cell cultures offer a convenient experimental model for studying evolutionary processes. Compared to more traditional genetic models (fruit flies and bacteria), the advantages are a relatively easy control and sampling of genotypes and of external conditions, short generation times, and, especially, high mutation rates (for RNA viruses). Application of the results presented in this paper and testing of their validity to these systems are rather straightforward. We list recommendations for two kinds of experiments.

Experiments on growth competition, aimed at comparing the fitness of two chosen genetic variants, are common in the virological literature (54–56, 58–60). They are typically carried out by mixing a majority of mutant virions with wild-type virus and monitoring the change in proportion as a function of repeated passage in permissive cells. The selective advantage(s) can then be estimated from the slope of the curve relating the mutant frequency, f, to the number of generations (Fig. 8). Note that the slope of the curve where it crosses 50% is independent of whether the experiment is carried out in the selection or the selection-drift regime. New spontaneous mutations may arise, changing the virus fitness and distorting results. This problem can be avoided if the population of infected cells is chosen in the selection-drift interval (Table 1), $1/s \ll N \ll 1/\mu$. Then, on the one hand, competition dynamics is almost deterministic, until one of the two subpopulations becomes very small. On the other hand, the time in which a mutation (advantageous for the virus but unwanted by the experimentalist) will appear, $1/(\mu Ns)$, is much longer than the measurement time, $1/s$, required to resolve the two growth rates (assuming that all selection coefficients are on the same order and that the advantageous mutant allele is not present in the initial population, i.e., a single ex vivo clone).

In the opposite experiment, one starts from a monomorphic population and monitors how fast a new advantageous mutation appears and outgrows the old genetic variant. To shorten the waiting time (Fig. 13), the population size must be large, at least in the selection-drift regime. After a new colony exceeds the critical size imposed by the stochastic bottleneck (see "Stochastic dynamics: the drift regime" above), the dynamics is almost deterministic.

Based on our results, the evolutionary experiment is not a suitable way to measure the spontaneous mutation rate. For example, attempted measurements of changes in the mutation rate in bacteria due to changes in external growth conditions (adaptive mutation) (61) are difficult to interpret. First, the selection coefficient is affected by the change in external conditions as well, and this effect is likely to be more important than the change in the mutation rate. As one can show (second equation 62), in a deterministically large population, even a substantial change in the mutation rate causes only a slight shift in the reversion curve (Fig. 8). Only the selection coefficient can be reliably assessed in such an experiment. Second, such experiments depend on the details of the evolutionary model. Third, if the population is small (Fig. 13), the time dependencies of the mutant frequency will fluctuate between different cultures and the changes in the mutation rate cannot be detected due to statistical error.

HIV populations in vivo. Data on the evolutionary behavior and genetic diversity of HIV, if understood in sufficient detail, could reveal vital information about major biological factors acting on the virus population in vivo. Of particular interest are

(a) $$\varepsilon^2 = \frac{2}{kT} - \frac{2(2k-3)}{k(k-1)}$$

(b) $$T_{rep} = \begin{cases} \frac{1}{lnN} & \text{"drift"} \\ \frac{1}{Nsln(1/s)} & \text{"selection-drift"} \\ \frac{2\mu}{s} & \text{"selection"} \end{cases}$$

(c)

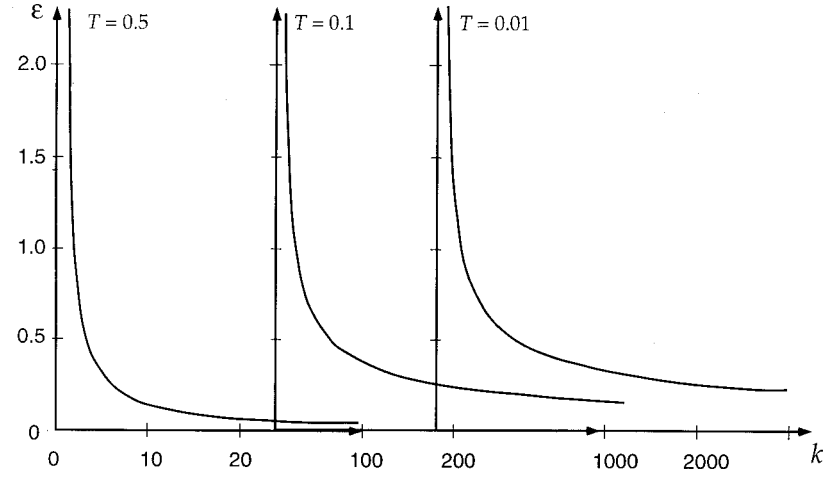

FIG. 15. Error in the measured genetic distance due to sampling effects. (a) The mathematical formula relating the standard error, ε, to the sample size, *k*, and to the actual genetic distance, *T*. (b) Representative values of the genetic distance for a polymorphic population in the three main intervals of the population size, *N*. (c) Plot of ε versus *k* at three values of *T*.

the relative roles of stochastic factors and selective forces, the role of purifying selection versus selection for diversity, and possible variation of wild-type sequence between individuals and tissues.

One application of this model is to use HIV genetic variation as a tool to probe the underlying size and structure of the infected cell population. There has been considerable controversy in the literature about the effective population size of HIV in a representative untreated patient. The concept of the effective size was introduced by Wright as a means of referring the intensity of genetic drift in a real population to that in an ideal Wright-Fisher population (80); i.e., the effective size of a real population is the size an ideal population would have if it also had the same rate of genetic drift as that observed in the real population. Another, perhaps more intuitive, way of thinking about the effective size is to consider the inbreeding effective population size, defined as the inverse of the probability that two randomly selected individuals have a common ancestor in the previous generation. (This is conceptually close to the crude "virological" definition of the effective population size as the number of productively effective cells that produce most of the virions that infect the next generation of productively infected cells.) If this probability of identity by descent is low, it stands to reason that the population size must be large, and vice versa. One begins to see how the (inbreeding) effective population size influences the genetic diversity. If the effective size is small, the probability of identity by descent is high, and there is consequently low genetic diversity, since individuals tend to be closely related. Both definitions apply in the presence of weak selection as well, e.g., for the model system shown in Fig. 1. There are several other measures of effective size, e.g., the variance effective population size and the eigenvalue effective population size, but these rarely give different values. Of course, the usefulness of all these definitions depends on the hypothesis that the actual population is not too far, in the sense of its evolutionary properties, from an ideal Wright-Fisher model with suitable parameters.

Assuming that selection is not important and the neutral model applies, a coalescence-based approach (see "Many loci and other aspects" below) has been used to estimate an effective size as small as 100 to 1,000 cells (45), much smaller than the total number of productively infected cells per patient, 10^7 to 10^9 (22). At least one other study reported similar values (66). However, other lines of evidence, including differences among rates of accumulation in different genes (74) and very high (44, 84) or very low (4, 5) ratios of synonymous to nonsynonymous mutations in some genes, imply that HIV populations are subject to significant selective influences. There-

fore, population genetic methods that assume a lack of selection may yield erroneous results.

Two of us recently developed and applied a robust method to estimate the effective HIV population size in vivo based on the genetic variation at close pairs of highly diverse sites (67). As follows from the simulation examples above (Fig. 8, 9, and 13), a site cannot preserve a high diversity ($f \simeq 0.5$) indefinitely. Early in infection, the HIV population is almost uniform genetically or has a limited number of sequences, due to the bottleneck that occurs at transmission and to early competition between clones (12, 27, 46, 86). Therefore, highly diverse sites are sites that are caught in the act of "reversion" from mutant to wild type (i.e., of advantageous substitution). The basis for this test was to select two such sites, A(a), and B(b), where the lowercase and capital letters denote mutant and wild type, respectively, and then classify all sequences in the population into four groups (haplotypes): ab, Ab, aB, and AB. During reversion, the population starts from an almost uniform haplotype ab and arrives at an almost uniform haplotype AB. The two other haplotypes are transient. The idea of the test is that, deep in a stochastic regime and given a limited sample size, one of the four haplotype groups will be empty at any time, because the time at which reversion ensues is random. Suppose that the population is deep in the selection-drift regime. Two sites revert typically at different random times, even if their selection coefficients are equal (Fig. 13). Nearly simultaneous reversion can happen accidentally. In all cases, as can be shown, the number of well-represented subclones (i.e., the number found in a sample of the usual size [10 to 30 clones]) is typically two, rarely three, and much more rarely four, at any time point.

Using sequence databases for HIV *pro* and *env* genes from drug-naive individuals (27, 44), we found that this effect is absent for close pairs of bases. We checked that this effect is not sensitive to variation of the initial genetic composition and some other factors assumed in the model and estimated the effect of recombination (derived from kinetic data) on the test to be numerically small as well. We therefore were able to conclude that a steady-state HIV population in an untreated individual, with respect to evolution of separate bases, is either in or at the border of the deterministic interval of population sizes.

Some authors considered a possibility that an HIV population may consist of weakly connected small populations. Shedding viruses from these subpopulations into the peripheral blood could explain the presence of all four haplotypes in the above test, in apparent contrast to our conclusion. Indeed, HIV-infected cells are located within lymphoid tissue in visually distinct islands (64). However, different islands exchange virions and infected cells and may or may not be weakly connected genetically. The strong overlap of the island patterns obtained for different virus strains (64) proves that the island structure is due to nonuniform distribution of infectible cells rather than to random seeding by the virus. Next, estimates based on studies of the clearance rate of free virus from peripheral blood (85), on HIV RNA quantitation in the lymphoid tissue (22), and on the decay rate of infectious virus titer under highly active antiretroviral therapy (62) suggest that a considerable portion of virus particles produced in the tissue drain into the blood (within a few hours or less, from where they are removed within a few minutes). This implies that a good portion of virus particles infect cells far from the cells that produced them, suggesting strong virus transfer between the islands within the same tissue. On the other hand, viruses isolated from some locations (semen and the central nervous system) show phylogeny distinct from that of the main virus reservoir in the body. Genetic sampling from different islands could clear the issue (28).

Given a relatively weak role of stochastic effects detected, we decided to test which factors shape evolution in the HIV protease (*pro*) gene (68). Using the same database of *pro* sequences, we observed that variation was restricted to rare bases: an average base was variable in about 16% of patients. The intrapatient distance per individual variable site, 27%, was similar for synonymous and nonsynonymous sites, although synonymous variable sites were twice as abundant, implying that purifying selection is the dominant kind of selection. We explained these facts within the one-locus model of evolution by assuming deterministic evolution within individuals and random sampling during the transmission between individuals. We considered different variants of the model with transmission of one and several genomes and with coinfection from independent sources. The model explained the variable sites as slightly deleterious mutants that are slowly being replaced with the better-fit variant during individual infection. In the case of a single-source transmission, genetic bottlenecks at the moment of transmission effectively suppress selection, allowing mutants to accumulate along the transmission chain to the high levels observed. However, we found that even very rare coinfections from independent sources are able to counteract the bottleneck effect and keep mutants at low levels. If such coinfections occur, the plausible explanation of the high level of mutants in an inoculum is variation of the best-fit sequence between individuals due to variation in the specific immune response, combined with coselection. In this model, variation in *pro* is due to a cascade of mutations compensating for early antigenic escape mutations. Note that our analysis was restricted to the single-locus approximation (see below).

General applications. The progress of evolution, in general, may be limited by the time it takes for a new advantageous (in our notation, wild-type) allele to appear and become fixed in the population. Figure 16 shows schematically the time required to reach a 50% composition as a function of the population size. An interesting conclusion about the relative role of selection and randomness follows from this diagram. The reversion half time depends on the population size. The shortest time, given mostly by the inverse selection coefficient, is reached in the deterministic limit, $\mu N >> 1$. This implies population sizes larger than 10^8 to 10^9 genomes for DNA viruses and bacteria and 10^{11} to 10^{12} for higher organisms. Such a population size exceeds the total size of any species higher on the evolutionary scale than insects. On the other hand, a mutation in a small population with a size smaller than the inverse selection coefficient (drift regime, in our terms) would be fixed only after a number of generations given by the inverse mutation rate, $1/\mu$, corresponding to timescales of planetary development. Put another way, nucleotides with a very small selective advantage compared to the inverse population size evolve very slowly. The characteristic values of selective advantage for the most important mutations (in an

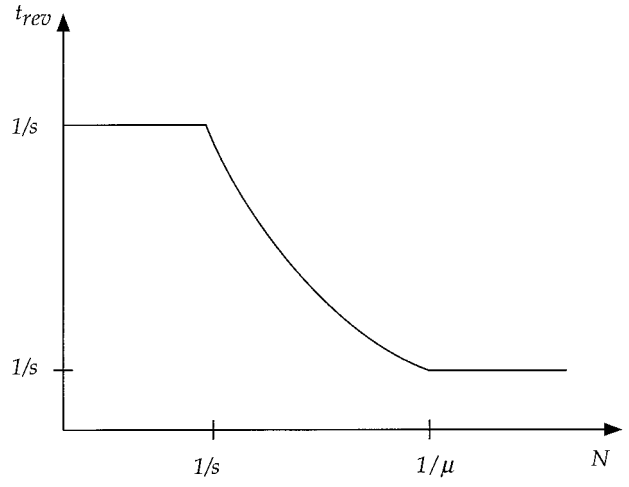

FIG. 16. Dependence of reversion time on population size. The schematic plot shows t_{rev}, the time required to reach 50% composition in the reversion experiment (fixation of advantageous allele), as a function of the population size, *N*.

evolutionary sense) are unknown. Still, the above considerations suggest the possibility that the evolution of higher organisms may be driven, mostly, by nucleotides with *s* larger than the inverse population size (provided that they are sufficiently frequent in the genome), i.e., within the selection-drift interval (Table 1). Then the rate of evolution does not have to be unreasonably low and the size of a population does not have to be unreasonably large. If this is the case, one can conclude that the two factors, random drift and selection, are equally important for the rate of evolution on very large timescales (millions of years) and that neither is a small correction. The reversion (fixation) half time within the interval $1/s < N < 1/\mu$ is $\sim 1/(\mu Ns)$.

Many Loci and Other Aspects

The distinction between deterministic and stochastic evolutionary genetic processes is critical—under a deterministic regime, the fate of a novel mutation or allele can be known with certainty; under a stochastic regime, we are able to characterize only the statistical properties of allele frequencies and fixation times. The neutral and the deterministic cases represent the ends of a spectrum, and precisely where a population sits depends most strongly, on its effective size. However, much of the theory relating effective population size and the stochastic-deterministic continuum had been worked out for simple models involving two (as in the present review) or, at most, a handful of alleles (37). These developments have allowed population geneticists to characterize the qualitative behavior of genes in populations undergoing a variety of dynamic processes, e.g., population size change, subdivision, and selection. Nonetheless, the assumptions and restrictions of these simple models typically preclude their use as descriptors of real populations, except in some particular cases (below).

In 1982, Kingman (39, 40) introduced a new way of studying the stochastic behavior of genes in a population. His framework—the coalescent—characterizes the genealogies of genes (or gene fragments), specifically the statistical distribution of times to common ancestors. It assumes the neutral model of evolution so that phylogenetic branching and accumulation of mutations are independent processes. Just as we estimated above (see "Decay of the polymorphic state and gene fixation"), the average time to the most recent common ancestor of a sample of genes is, within a numerical factor which depends on the sample size, the effective population size. The coalescent incorporates the same information as allele-based methods but in a form more relevant for today's evolutionary geneticist. This is largely because the raw data of molecular evolutionary studies in the last 15 years have been the molecular sequences, and there is an extensive and well-developed literature on molecular phylogenetic reconstruction (reference 65 and references therein). Studies on the coalescent have also shown that there is an increase in the power of parameter estimation as more independent loci are analyzed. Selection—for a long time the thorn in the side of the coalescent theorist—can, in principle, be accommodated within a coalescent-like framework. The recently developed inclusion of selection in a genealogical framework (42, 55) requires the construction of an ancestral selection graph, akin to a coalescent genealogy, which has the unusual property of coalescing and splitting as one moves back in time. The reader has to keep in mind, though, that a mathematical theory on a network (the case with selection) is technically much harder to handle (with or without computer simulation) than a theory made for the tree topology as the neutral coalescent. The tree theory, but not the network theory, can be reduced to one-dimensional chain of equations. (Similar issues arise in physics, in theoretical studies of hopping transport of electrons [63, 72].) Still, this recent achievement should stimulate the development of novel methods that work with selection as well.

In some cases, recombination or point mutations break linkage disequilibrium and make loci almost independent, so that the one-locus approximation applies directly. One such case is when strong recombination is present in the system and the variable loci are spread far apart (for HIV, the respective length is predicted to be around 100 nucleotides or longer, if superinfection protection is efficient [67]). If the recombination rate is low, Muller's ratchet (6, 13, 14, 17, 53) may operate. With Muller's ratchet, random drift can lead to the elimination of the fittest genomes from a population. Once a better-fit haplotype is accidentally lost from a recombination-free population, it cannot reappear, thus clicking the ratchet another notch. Successive "clicks" cause the population to become successively less fit, on average. Back mutations can, in principle, restore the disappearing better-fit haplotypes. For a long segment of genome of a replication-competent virus, they are expected to be much less likely than the forward mutations, since the frequency of deleterious substitutions is expected to be low. As follows from Monte Carlo simulation, which we hope to discuss elsewhere, back mutations can prevent Muller's ratchet only for sites with large selection coefficients, $s = 0.1$ to 0.2, and provided that the gene segment is short.

Another important consideration is that selective pressures on some parts of the genome must also have an effect elsewhere. Such "background" selection may explain the very small effective size of HIV obtained from the *env* gene diversity under the neutral approximation. Charlesworth and colleagues (7, 8, 57) have shown that if an unsequenced region of a

genome is under selection and is linked to the region under study, depressed estimates of effective size are obtained. Potentially, then, selective pressures on *gag* or *pol* can influence diversity in *env*. However, linkage disequilibrium obviously has an effect beyond simply lowering the effective size, and it speaks to the issue of fitness at the unit of the individual virion. If there is linkage disequilibrium, does it really make sense to look for the effects of immune-driven selection only in *env* or *gag*, as many studies have done (3, 70, 83)? To what extent is the fitness advantage of mutations in *env*, for instance, balanced by the loss of fitness due to mutations in *pol*? Such "interaction" between loci in a small population may be caused by linkage alone and applies even to mutations that additively affect the fitness of a genome. To make things more complex, nonadditive compensatory mutations (epistasis) exist, due to actual biological interaction between loci, at both the nucleotide and protein levels. Moreover, in vesicular stomatitis virus systems, epistasis may be the factor counteracting the loss of fitness due to Muller's ratchet (16). Compensatory mutations, which become advantageous only after initial resistance mutations occur, have also observed in HIV-infected patients treated with protease inhibitors (11). The development of molecular techniques that allow full-length HIV genomes to be sequenced (69) means that it is only a matter of time before genetic data become available to study the evolutionary processes of linkage disequilibrium, background selection, and compensation in infected individuals.

Conclusions

We have analyzed in depth a broad range of problems in evolutionary dynamics in the framework of a simple one-locus, two-allele population model, which includes three basic factors: random drift, point mutation, and selection. We found that (as long as the mutation rate is lower than the selection coefficient) the dynamic properties differ drastically in three wide intervals of the population size that we call the drift, selection-drift, and selection regimes. Transition between stochastic and deterministic behavior of genetic evolution occurs in the intermediate selection-drift regime, which is expected to be very wide in the population size, especially for DNA systems. In this regime, deterministic laws govern genetically highly polymorphic populations, and almost uniform populations evolve stochastically.

Estimates of typical population sizes and of the time in which new advantageous alleles appear and become fixed in the population suggest that higher organisms may evolve while in the selection-drift regime. If this is the case, the speed of evolution depends on three parameters: mutation rate, selective advantage, and population size. Hence, selection pressure and random drift, whose relative importance for evolution is often disputed in the literature, are equally important, although they act differently: selection promotes evolution, and random drift slows it down.

The theory provides recommendations for the size of the population in different bacteriological and virological experiments in vitro aimed at either comparing the fitness of different mutants or measuring the mutation rate. For HIV populations in vivo, theory based on the purifying selection alone predicts either a weak diversity or a very low genetic turnover rate. Experimental searches for rapidly varying bases can provide biological evidence for selection for diversity due to different environments, a changing immune response, changes in host cell populations with time, and other important aspects of HIV infection.

Naturally, with any research program that requires theory to be integrated with data, there is an inevitable tension between experimental biologists, who deal daily with the complexity of real biological systems, and theoretical biologists, who "simplify, simplify, simplify" in the name of tractability. In this work, our analysis has been limited to the simplest possible case: evolution of a single locus with only two alleles. Many important aspects of evolution, including the effects of multiple loci, recombination, coselection, and migration, were not considered. Nevertheless, in-depth consideration of this simple system has yielded a surprisingly rich set of results, which should be very useful for the design of experiments in evolution and for the interpretation of patterns of genetic variation in natural infection. In the future, however, we see greater reliance on the fusion of analytic and computational methods as a means of simulating the complexity of real populations. By tying these computer-intensive methods to well-characterized mathematical and statistical methods, one has the advantage of using standard inferential procedures without sacrificing too much in the way of realism. However, the old adage that one has to walk before one can run applies to population genetics as it does elsewhere, and understanding simple evolutionary models is perhaps the surest route to coming to grips with the complexity of virus evolution.

MATHEMATICAL RESULTS AND DERIVATIONS

Description of the Model and the Evolution Equation

In this section, we will derive the diffusion-type differential equation and complement it with the boundary conditions. First, we derive the discrete Markovian equation for the virus population model; second, we reduce it to the continuous diffusion equation; and third, we determine the boundary conditions for the diffusion equation, in different intervals of the population size. In other sections, we will solve the appropriate set of equations and boundary conditions for each interval of N and different initial conditions. Table 2 contains a list of the principal notation used in this section.

Main results. We show that the stochastic evolution of the virus population is described by one of two different sets of differential equations and boundary conditions, depending on the interval of the population size, N. A large population, $\mu N \gg 1/\ln N$, usually has many copies of both the wild-type and mutant genomes. The corresponding evolution equation and the boundary conditions have the form (32)

$$\frac{\partial \rho}{\partial t} = -\frac{\partial q}{\partial f} \tag{1}$$

$$q(f,t) = -\frac{1}{2N}\frac{\partial}{\partial f}[f(1-f)\rho] - sf(1-f)\rho - \mu(2f-1)\rho \tag{2}$$

$$q(f,t)_{f\to 0} = q(f,t)_{f\to 1} = 0 \tag{3}$$

where f is the mutant frequency, ρ is the probability density, t

TABLE 2. Mathematical notation

Symbol	Definition
A, B, C, F	Undetermined constants or functions
D	Relative interpopulation distance per site
$\delta(x)$	Dirac delta function of *x*
δ_{ij}	Kroneker symbol: 1 if $i = j$ and 0 otherwise
f	Mutant frequency
G	Gene fixation probability
g	Continuous part of the probability density
K	Time correlation function
μ	Mutation rate per generation per site
M	Mean change in the mutant frequency per unit time
N	Population size (productively infected cell number)
n	Mutant genome number
P_n	Probability of having *n* mutant genomes
p_0	Probability of having a pure wild-type population
p_1	Probability of having a pure mutant population
q	Probability density flux
ρ	Probability density of the mutant frequency
s	Selection coefficient
T	Intrapatient genetic distance per site
t	Time (generation number)
V_x	Variance of *x*
x	Any parameter (for this table only)
$\bar{x}$	Expectation value of *x*
x_{ss}, x^{ss}	Value of *x* in steady state

is the generation number (time), and *s* is the selection coefficient. Equations 1 to 3 are valid under the conditions $s \ll 1$, $\mu \ll 1$, and $\mu N \gg 1/\ln N$. In this case, *t* and *f* can be treated (approximately) as continuous variables. Effects of the three terms in the right-hand side of the evolution equation given by equations 1 and 2 are illustrated in Fig. 2.

A useful analogy between the probability density and the density of a gas between two walls is discussed in the qualitative section of this review (Fig. 3). In this analogy, equation 1 expresses the fact that gas particles do not appear or disappear but only travel from one location to another. The quantity *q(f,t)* in equation 2 is the "probability flux," analogous to the gas flux density defined as the net number of particles crossing a plane at *f* from left to right per unit area per unit time. Thus, the evolution equation written in the form of equation 1 expresses the fact that the probability density is a locally conserved quantity, just like a gas density. The boundary conditions in equation 3 state that the probability flux vanishes at the boundaries of the allowed interval in *f*, similar to gas particles being prohibited from crossing the confining walls (Fig. 3b).

In small populations, where $N \ll 1/[\mu\ln(1/\mu)]$ (see below), the population can be found, with a finite probability, in a purely monomorphic state, of $f = 0$ or 1 (similar to condensation of gas at cold walls). In this interval, we break up the total probability density into a sum of the continuous probability density and of two singular terms, as given by

$$\rho(f,t) = p_0(t)\delta(f) + p_1(t)\delta(1-f) + g(f,t) \quad (4)$$

where p_0 and p_1 are the probabilities of having pure wild-type and pure mutant, respectively; $\delta(f)$ denotes the Dirac delta function; and $g(f,t)$, where $f(1-f) \gg 1/N$, is the continuous part of the probability density. The boundary conditions are

$$\frac{dp_0}{dt} = -q(0,t),\ \frac{dp_1}{dt} = q(1,t),\ N \ll 1/[\mu\ln(1/\mu)] \quad (5)$$

$$2\mu N p_0 = [fg(f)]_{f\to 0},\ 2\mu N p_1 = [(1-f)g(f)]_{f\to 1} \quad (6)$$

The first pair of conditions (equations 5) describes the accumulation or depletion of probability at the two boundaries, analogous to condensation or evaporation of gas (Fig. 3c). The second pair (equations 6) reflects the fact that transition between a monomorphic state, $f = 0$ or 1, and the closest polymorphic state, $f = 1/N$ or $(N-1)/N$, respectively, can occur due to a mutation only. This pair of equations has to be derived from the first principles, i.e., the discrete virus population model (below). The differential equation for the continuous part of the probability density, *g(f,t)*, has a form

$$\frac{\partial\rho}{\partial t} = -\frac{\partial q}{\partial f},\ q(f,t) = -\frac{1}{2N}\frac{\partial}{\partial f}[f(1-f)\rho] - sf(1-f)\rho \quad (7)$$

which differs from the expressions at large *N* (equations 1 and 2), in that the term with μ in equation 2 is absent. The mutation rate enters the problem through the boundary conditions (equation 6) only.

One can easily obtain the upper bound for *N*, within which the above boundary conditions apply, from the boundary conditions themselves. The probability of a polymorphic state is given by $\int_0^1 g(f)df$. As follows from equation 6, near the boundaries *g(f)* diverges and is given by $g(f) \sim 2\mu N p_0/f$ and $g(f) \sim 2\mu N p_1/(1-f)$. The integral of *g(f)* is mostly contributed from $f \simeq 1$ or 0 and is truncated at $f(1-f) \sim 1/N$. The resulting probability of a polymorphic state is comparable to the probability of a monomorphic state, $p_0 + p_1$, at $\mu N \log N \simeq 1$, as we stated above.

Note that the validity of these equations, as we discussed in the qualitative section of this review, is not restricted to the virus population model. The same diffusion equation (equation 1 and 2) applies to many other haploid one-locus, two-allele populations, which include the same three factors: random sampling of genomes, symmetric mutations, and purifying selection. Should some other factors come into consideration, such as allelic dominance in a diploid population or time fluctuations of selection coefficient or of other parameters, a more general equation of similar form can be written (see "stochastic equation of evolution" below) (33). In principle, the approach can be generalized for many-allele or multiple loci using partial derivatives in haplotype frequencies (37).

Now we proceed with derivations of all these formulas.

Virus population model. The model (as discussed in the qualitative section of this review) considers an asexual population of *N* cells infected with two genetic variants of a virus: *n* cells are infected with "mutant" virus, and $N - n$ cells are infected with "wild-type" virus. The total population size *N* is fixed, while *n* changes in time. During a generation step, each mutant-infected cell produces b_1 mutant virions and then dies, and each wild-type-infected cell produces b_2 wild-type virions and dies (Fig. 1b). The respective numbers of virions per cell, b_1 and b_2, are assumed to be large, $b_1 \gg 1$, $b_2 \gg 1$, and differ slightly for the two alleles

$$b_1 = b_2(1-s) \quad (8)$$

where s, $s \ll 1$, is, by definition, the selection coefficient (mutation cost), reflecting the difference in fitness. From all the virions produced per generation, N virions are sampled randomly to infect new generation of cells. Each virion, on infecting a cell, can mutate into the opposite genetic variant with a probability μ, $\mu \ll 1$. The virus population model described is a particular case of the Wright-Fisher population with discrete time.

Stochastic equation of evolution. (i) Discrete Markovian equation. Let $p(n,t)$ be the probability of n mutant cells at time t, where t is an integer that numbers generations and n can change from 0 through N. If consecutive generations do not overlap, $p(n,t)$ is a Markovian process described by a discrete evolution equation

$$p(n,t+1) = \sum_{n'=0}^{N} P(n|n')p(n',t) \tag{9}$$

where $P(n|n')$ is the conditional probability of having n mutants, given that their number at the previous step was n'. In this section, we derive $P(n|n')$ for the virus population model introduced in the qualitative section of this review.

First, we obtain the conditional probability $P(n|n')$ in equation 9, neglecting mutation events and denoting the result $P_0(n|n')$. Suppose that the number of mutants in some generation is n'. The total numbers of virions produced by all mutant- and all wild-type-infected cells are

$$B_1 = b_1 n',\ B_2 = b_2(N - n') \tag{10}$$

respectively. A biologically reasonable assumption is $b_1, b_2 >> 1$ (above). If n is the number of new mutant-infected cells, then the numbers of mutant and wild-type virions which infect must be n and $N - n$ virions, respectively. The probability of n new mutant cells, $P_0(n|n')$, is proportional to the number of possible ways in which one can choose n mutant virions from B_1 possible mutant virions and $N - n$ wild-type virions from B_2 possible wild-type virions

$$P_0(n|n') = A\,\frac{(n')^n(N-n')^{N-n}(1-s)^n}{n!(N-n)!} \tag{11}$$

where we used equations 8 and 10 and the constant A is determined by the condition $\Sigma_n P_0(n|n') = 1$.

We now take mutations into consideration. Suppose, at the moment of infection of new cells by n mutant and $N - n$ wild-type virions, m_1 forward and m_2 reverse mutations occur (Fig. 1b). The resulting number of mutant-infected cells, n'', will be $n'' = n + m_1 - m_2$. The probability of m_2 reverse mutations among n infecting virions, if n is large, is given by Poisson statistics with the average μn

$$\pi(m_2|n) = \frac{(\mu n)^{m_2}}{m_2!}\,e^{-\mu n},\ m_2 = 0,1,\ldots \tag{12}$$

(If n is not large, equation 12 still is valid for $m_2 = 0$ and 1, which are the only important values in this case, since we assume $\mu \ll 1$ everywhere in the present work.) Analogously, the probability of m_1 forward mutations is $\pi(m_1|N-n)$. As a result, for the conditional probability $P(n''|n')$, we obtain

$$P(n''|n') = \sum_{n=0}^{N}\sum_{m_1=0}^{N-n}\sum_{m_2=0}^{n}\delta_{n'',n+m_1-m_2}\pi(m_1|N-n)\pi(m_2|n)P_0(n|n'). \tag{13}$$

where $P_0(n|n')$ is given by equation 11 and the Kroneker symbol $\delta_{i,j}$ is 1 if $i = j$ and 0 otherwise.

(ii) Diffusion equation limit. The discrete evolution equation given by equations 9, 11, and 13 may be suitable for computer simulation, but is very inconvenient for analytic treatment. In addition, it contains many model-dependent details which are not important at long timescales and in large populations and could be best disregarded. When both subpopulations are large ($n \gg 1$ and $N - n \gg 1$), equation 9 can be transformed to a differential form. In this case, the conditional probability, $P(n|n')$, changes slowly in n and n', $|P(n|n') - P(n+1|n')| \ll P(n|n')$, $|P(n|n'+1) - P(n|n')| \ll P(n|n')$, and can be approximated by a continuous function of n and n'. Substituting the asymptotic formula, $n! \simeq (2\pi n)^{1/2}(n/e)^n$, into equation 11, we find that $P_0(n|n')$ has a narrow maximum at $n' \simeq n$. Rewriting $P_0(n|n') = \exp[\ln(P_0(n|n'))]$ and expanding the argument of the exponential in n' up to the second-order terms in $n' - n$, we obtain

$$P_0(n|n') = A\,\exp\left\{-\frac{[n-n'+sn'(1-n'/N)]^2}{2n'(1-n'/N)}\right\} \tag{14}$$

Note that for the characteristic half-width in equation 14 we have $1 \ll |n - n'| \ll \min(n', N - n')$, which confirms that $P_0(n|n')$ can be considered a smooth function of its variables.

Note that the style of the last paragraph is the approximate derivation "with a large parameter," standard for theoretical physics. In this case, large parameters are n, N, $1/s$, and $1/\mu$. One makes a guess that the function $p(n|n')$ is smooth and verifies its consistency later, after the result was obtained. To avoid a vicious circle, one checks alternative assumptions as well. A more rigorous way, usual in population biology, would be to evaluate probability $p(n, n')$ in the limit $N \to \infty$, $\mu \to 0$, $s \to 0$ while μN and sN are constant. In our experience, the method we employ almost always leads to a correct result and is easier to use, especially when there exists an independent verification of the result.

We now obtain a simplified expression for the full conditional probability in equation 13 using the fact that mutations are rare ($\mu \ll 1$), so that the probable values of m_1 and m_2 are much smaller than those of $N - n$ and n. We substitute $n'' - m_1 + m_2$ for n in the arguments of both π functions and of the function P_0 in the right-hand side of equation 13 and expand these functions in $m_1 - m_2$, up to the first-order terms. The resulting double sum in m_1 and m_2 can be evaluated exactly with the use of equation 12, which yields, within the same accuracy,

$$P(n|n') = (1+2\mu)\,P_0(n|n') + \mu(2n-N)\frac{\partial}{\partial n}P(n|n') \simeq P_0(n+\mu(2n'-N)|n') \tag{15}$$

Since the probability $p(n, t)$, with one exception discussed below separately, is a smooth function of n, it will be more

convenient, from now on, to consider the probability density $p(f,t)$ of the mutant frequency $f \equiv n/N$, normalized by the usual condition $\int_0^1 df\ \rho(f,t) = 1$.The two definitions are related, as given by $\rho(f,t) = Np(Nf,t)$. In terms of $\rho(f,t)$, the evolution equation, given by equations 9, 14, and 15, can be written as

$$\rho(f,t+\epsilon) = \int df' \Pi_\epsilon(f|f')\rho(f',t) \tag{16}$$

$$\Pi_\epsilon(f|f') = (2\pi\epsilon V(f'))^{-1/2} \exp\left[-\frac{(f-f'-\epsilon M(f'))^2}{2\epsilon V(f')}\right] \tag{17}$$

$$M(f) = -sf(1-f) - \mu(2f-1) \tag{18}$$

$$V(f) = \frac{1}{N} f(1-f) \tag{19}$$

where $\epsilon = 1$ is the time between generations. In the continuous approximation, ϵ in the above expressions can be replaced by any small time interval. The notations $M(f)$ and $V(f)$, introduced in equations 18 and 19, have meanings of the expectation value and of the variance of change in f per unit time, respectively. One can present them in the general form

$$M(f') = \epsilon^{-1} \int df (f-f') \Pi_\epsilon(f|f'), \tag{20}$$

$$V(f') = \epsilon^{-1} \int df (f-f'-\epsilon M(f'))^2 \Pi_\epsilon(f|f') \tag{21}$$

The validity of the latter relationships can be checked substituting equation 17 into equations 20 and 21.

As shown below, the integral equation, (equations 16 and 17) can be transformed to the Kolmogorov forward equation

$$\frac{\partial \rho}{\partial t} = \frac{1}{2}\frac{\partial^2}{\partial f^2}(V\rho) - \frac{\partial}{\partial f}(M\rho) \tag{22}$$

which, together with equations 18 and 19, yields the promised diffusion equation, equations 1 and 2.

We will derive equation 22 from equations 16 and 17 in a general form, without specifying the model-dependent parameters M and V from equations 1 and 2. We will assume that $V(f) \ll 1$ and that the higher momenta $\overline{(f-f')^3}, \overline{(f-f')^4}, \ldots$ of the conditional probablity $\Pi_\epsilon(f|f')$ are proportional to powers of ϵ higher than 1. These assumptions are valid for our model, as can be checked using equations 17 to 19 (see "Virus population model" above). The derivation that follows is well known.

Consider an arbitrary function, $A(f)$, localized in the interval $0 < f < 1$ far from its ends, so that $A(f)$ and its derivative vanish at $f = 0$ and $f = 1$. We also introduce the expectation value

$$\overline{A}(t) = \int df\, A(f)\rho(f,t) \tag{23}$$

Multiplying both sides of equation 16 by $A(f)$ and integrating in f, we have

$$\overline{A}(t+\epsilon) = \int df' \rho(f',t) \int df\, A(f) \Pi_\epsilon(f|f') \tag{24}$$

Since the characteristic width of $\Pi_\epsilon(f|f')$ in terms of $f - f'$ is small, we are allowed to expand $A(f)$ in the integrand in equation 24 in a series of $f - f'$. Evaluating the resulting integral over f and discarding terms of higher than first order in ϵ, we get

$$\overline{A}(t+\epsilon) = \overline{A}(t) + \epsilon \int df' \rho(f',t) \left[\frac{1}{2} V(f') \frac{d^2A(f')}{df'^2} + M(f') \frac{dA(f')}{df'} \right] \tag{25}$$

where we used the definitions in equations 20 and 21. Higher-than-second terms of expansion of $A(f')$ can be neglected due to the above assumption. Evaluating the integral over f' in equation 25 by parts and using the definition of the time derivative, we get

$$\frac{d\overline{A}}{dt} = \int df'\, A(f') \left[\frac{1}{2} \frac{\partial^2}{\partial f'^2} \Big(V(f')\rho(f') \Big) - \frac{\partial}{\partial f'} \Big(M(f')\rho(f') \Big) \right] \tag{26}$$

We arrive at the desired evolution equation (equation 22) by choosing $A(f') = \delta(f' - f)$. Here the width of the "delta function" is assumed to be much larger than $V(f)$ but much smaller than the characteristic values of f at which the density function $\rho(f)$ changes noticeably.

Boundary conditions: properties of an almost monomorphic population. In a large population, $N \gg 1/\mu$, the boundary conditions have a form 3, which states that the probability density flux $q(f,t)$, defined as such by equation 1, must vanish at the boundaries, $f = 0$ and $f = 1$. This follows from the continuity conditions at the boundaries and from the understanding that monomorphic states, in which f is exactly 0 or exactly 1, are very unlikely to occur when the population is large due to a high mutation rate per population ($N\mu \gg 1$). As we show now, the flux does not vanish at the boundaries in smaller (but still very large) systems.

Suppose, first, that boundary conditions in equation 3 do apply. As follows from equations 2 and 3, the function $\rho(f,t)$ diverges near the boundaries at $\mu N < 1/2$. Indeed, solving the equation $q(f,t) \equiv 0$ near $f = 0$ and near $f = 1$, one obtains

$$\rho(f,t) \simeq \begin{cases} C_0 f^{2\mu N - 1}, & f \ll 1 \\ C_1 (1-f)^{2\mu N - 1}, & 1 - f \ll 1 \end{cases} \tag{27}$$

where C_0 and C_1 are constants. Integrating the first of equations 27 from $f = 0$ to $f \sim 1/2$ and the second from $f \sim 1/2$ to $f = 1$, one finds that the region of f, $1 - f$, such that $\ln(1/f)$ or $\ln[1/(1-f)] \sim 1/\mu N$, contributes most to the normalization integral. If the population is not too small, $N\mu \ln N \gg 1$, these values of f correspond to many copies of a minority allele, f, $1 - f \gg 1/N$. Therefore, for the probability of monomorphic states we have $p_0, p_1 \ll 1$. The boundary conditions given by equation 3 apply. If, however, the population is small, $N\mu \ln N \ll 1$, the most probable values of f are in the region f, $1 - f \ll 1/N$, corresponding to much less than one minority copy per entire population. The above result indicates that the population can be found, with a finite and even high (close to 1)

probability, in a state in which f is exactly 0 or 1. To account for this fact, we separate singular terms in $\rho(f)$, as given by equation 4, and obtain new boundary conditions. Since we now have two more time-dependent variables, P_0 and p_1, unlike in the case with a large N, we will need four rather than two conditions at the boundaries. The first pair of equations (equations 5) describe the continuity condition at the boundaries. We now derive the second pair.

We return to the discrete probability notation, $p(n)$. It suffices to consider only one of the boundary regions in n, say, $n \ll N$; the conditions for the other region, $N - n \ll 1$, are analogous. Therefore, the probability has two components: a large value $p(0,t) \equiv p_0(t)$ and a relatively small part $p(n,t)$, $n \neq 0$, which changes slowly with n at $n \gg 1$. [Strictly speaking, $p(n,t)$ is diverging as $1/n$ as $n \to 0$ (equation 27). Divergence of the integral $\int_n p(n,t)dn$, however, is only logarithmic at best, which is sufficiently slow for what follows.]

We start by simplifying equations 11 and 13 for $P_0(n|n')$ and $P(n|n')$. Using the condition $n \ll N$, equation 11 gives

$$P_0(n|n') \simeq \frac{[n'(1-s)]^n}{n!} e^{-n'(1-s)} \tag{28}$$

The same inequality ($n \ll N$) allows one to neglect the reverse mutations, keeping only terms with $m_2 = 0$ in equation 13. Next, the condition of small μN means that even a single forward mutation per generation per entire population is a rare event. Hence, one can discard in equation 13 all terms with $m_1 \geq 1$, except the term with $m_1 = 1$ for particular values $n'' = 1$, $n' = 0$. The latter term has to be kept since the transition between states with $n' = 0$ and $n'' = 1$ cannot occur by means other than a mutation. As a result, the expression for $P(n|n')$ simplifies to

$$P(n|n') = (1 - \mu N)P_0(n|n') + \mu N \delta_{n,1}\delta_{n',0} \tag{29}$$

We now introduce the characteristic polynomial of the probability function $\varphi(x,t)$

$$\varphi(x,t) = \sum_{n=0}^{N} p(n,t)(1-x)^n \equiv p_0(t) + \phi(x,t) \tag{30}$$

where $0 < x < 1$, and $\phi(x,t)$ is a sum over the continuous part of $p(n,t)$ for $n \neq 0$ only. The evolution equation for $\varphi(x,t)$, which can be obtained from equations 9 and 28 to 30, has a form

$$\varphi(x,t+1) = (1 - \mu N)\varphi[1 - e^{-x(1-s)}, t] + \mu N(1-x)p_0(t) \tag{31}$$

The characteristic values of n in $p(n,t)$ for $n \neq 0$ are large ($n \gg 1$). Hence, as one can see from equation 30, the characteristic scale of x for function $\varphi(x,t)$ is small ($x \ll 1$). We expand the right-hand side of equation 31 to the second-order terms in small x, which yields

$$\frac{dp_0}{dt} + \frac{\partial \phi}{\partial t} = -\left(sx + \frac{x^2}{2}\right)\frac{\partial \phi}{\partial x} - x\mu N p_0 \tag{32}$$

Here we substituted $\varphi = p_0 + \phi$ and made use of the strong inequalities $s \ll 1$, $\mu N \ll 1$, and $\phi \ll p_0$. We notice that at $x \ll 1$, the function $\phi(x,t)$ represents the Laplace transform

$$\phi(x,t) = \int_{0+}^{\infty} dn\, e^{-xn} p(n,t) \equiv \mathcal{L}_x\{p(n,t)\} \tag{33}$$

Using the operator of the Laplace transform $\mathcal{L}_x$, one can rewrite equation 32 in the form

$$\mathcal{L}_x\left\{\frac{\partial p(n,t)}{\partial t} + \frac{\partial q(n,t)}{\partial n}\right\} = \left[-q(n,t)_{n\to 0} - \frac{dp_0}{dt}\right] + \frac{x}{2}\{[np(n,t)]_{n\to 0} - 2\mu N p_0\} \tag{34}$$

where $q(n,t)$ is the probability flux

$$q(n,t) = -\frac{1}{2}\frac{\partial (np)}{\partial n} - snp \tag{35}$$

which coincides with definition of q in equation 7 at $f \equiv n/N \ll 1$. It is important that the probability function, $p(n,t)$, together with its derivatives at $n \neq 0$, is a nonsingular function of n, meaning that it does not contain a delta function or its derivatives. The singular part of the probability function is already separated in the term p_0. As is well known, a Laplace transform $\mathcal{L}_x$ of a nonsingular function cannot be constant and cannot increase with x in the limit of large x. Therefore, both bracketed terms in equation 34 must be zero, and we arrive at the boundary conditions at $f \to 0$ (equations 5 and 6). Since the left-hand side of equation 34 turns out to be zero, the argument of the Laplace operator in braces is zero as well. This yields the differential evolution equation (equation 7) at $f \ll 1$. The boundary conditions at another boundary, $f \to 1$ ($n \to N$), are obtained in the same manner.

Experiments on Evolution and Observable Parameters

In this section, we define rigorously the observable parameters introduced in the qualitative section of this review.

Let parameter $A(f)$ be a deterministic function of the random mutant frequency f. The expectation value of A, $\bar{A}$, and the variance, V_A, are defined by

$$\bar{A}(t) \equiv \int_0^1 df\, A(f)\rho(f,t) \tag{36}$$

$$V_A(t) \equiv \overline{(A - \bar{A})^2} = \bar{A}^2 - (\bar{A})^2 \tag{37}$$

V_A is equal to the square of the standard deviation of parameter A.

The intrapopulation genetic distance

$$T = 2f(1-f) \tag{38}$$

is the probability that a pair of sequences randomly sampled from a population with composition f differ at a given nucleotide; it varies between 0 and 0.5 (Nei's nucleotide diversity measure). The expectation values of parameters f and T are

given by equations 36 and 37, with $A(f) = f$ and $A(f) = 2f(1 - f)$, respectively. The variance V_f and the average $\bar{T}$ are related:

$$V_f = \bar{f}(1 - \bar{f}) - \bar{T}/2 \quad (39)$$

We also introduce the interpopulation genetic distance, T_{12}, which is defined as the probability that a pair of genomes sampled from two different populations with mutant frequencies f_1 and f_2 differ at a given nucleotide, and the relative distance between populations, D, as follows:

$$T_{12} = f_1(1 - f_2) + f_2(1 - f_1) \quad (40)$$

$$D = T_{12} - (T_1 + T_2)/2 = (f_1 - f_2)^2 \quad (41)$$

If the two populations are statistically independent, one has $\bar{D} = (\bar{f}_1 - \bar{f}_2)^2 + V_{f1} + V_{f_2}$. Other definitions for the genetic distance between populations have been proposed in the literature (see reference 54 and references therein).

Consider now the genetic divergence experiment. Suppose that two populations have been isolated, at $t = 0$, from the same parental population, which was at steady state. New populations are then allowed to grow quickly to the original size, so that their composition does not change from the (random) composition of the parental population ($f = f_0$). Our aim is to monitor how the average relative distance between population, $\overline{D}$, evolves after the moment of split. The expectation value, $\overline{D}$, is given by

$$\overline{D} = \int_0^1 df_1 \int_0^1 df_2 \int_0^1 df_0 (f_1 - f_2)^2 \, \rho(f_1,t|f_0)\rho(f_2,t|f_0)\rho_{ss}(f_0) \quad (42)$$

where $\rho_{ss}(f_0)$ denotes the steady-state probability density and $\rho(f,t|f_0)$ is the probability density, which satisfies the initial condition

$$\rho(f,0|f_0) = \delta(f - f_0) \quad (43)$$

Equation 42 can be also written in the form

$$\overline{D}(t) = 2\int_0^1 df_0 \, V_f(t|f_0)\rho_{ss}(f_0) \quad (44)$$

where $V_f(t|f_0)$ is the variance of f, defined by equations 36 and 37, with $A(f) = f$ and under the initial condition of equation 43. The variance $V_f(t|f_0)$ varies from 0 at $t = 0$ to its equilibrium value V_f^{ss} at $t = \infty$. Note that equation 44 reduces the relative distance between two population to the properties of one population.

Consider now a single steady-state population. Random variation of f in time can be characterized by the time correlation function

$$K(t) = \frac{1}{V_f^{ss}} [\overline{f(t)f(0)} - (\bar{f}_{eq})^2] \quad (45)$$

The choice of the initial moment $t = 0$ in equation 45 is arbitrary since the system is at steady state. The function $K(t)$ varies from 1 at $t = 0$ to 0 at $t = \infty$. The characteristic timescale of the random fluctuations, t_{trn} (the genetic turnover time) is defined by $K(t_{trn}) = 1/e$. The correlation function $K(t)$ can also be expressed in terms of the expectation value $\bar{f}(t|f_0)$, as given by

$$K(t) = \frac{1}{V_f^{ss}} \int_0^1 df_0 [\bar{f}(t|f_0) f_0 \rho_{ss}(f_0)] - (\bar{f}_{ss})_2 \quad (46)$$

where $\bar{f}(t|f_0)$ is defined by equation 36 with $A(f) = f$ and under the initial condition given by equation 43.

Steady State

In this section, we derive the general expression for the probability density in the steady state and discuss in detail the crossover between stochastic and deterministic behavior in two cases: in the neutral case ($s \ll \mu$) and in the opposite limit ($s \gg \mu$).

General case. From equations 1 and 3, which apply at large population sizes, $N\mu >> 1/\ln N$, and the condition of steady state, $\partial\rho/\partial t = 0$, we obtain

$$q(f) \equiv 0 \quad (47)$$

with $q(f)$ given by equation 2. Separating the variables f and p in the obtained differential equation and integrating both sides, we get (80)

$$\rho_{ss}(f) = C[f(1 - f)]^{-1+2\mu N} e^{-2Nsf}, \; N\mu \gg \frac{1}{\ln(1/\mu)} \quad (48)$$

where C is a normalization constant.

As discussed above in the section on boundary conditions, at small N such that $N\mu \ll 1/\ln N$, the probability density has singular components, equation 4, and obeys equations 5 to 7. From the steady-state conditions, $\partial g/\partial t = 0$ and $dp_0/dt = dp_1 = 0$, we obtain, again equation 47, with $q(f)$ given by equation 2 with $\mu = 0$, which yields

$$\rho_{ss}(f) = g_{ss}(f) + \frac{1 - p_{pol}}{1 + e^{-2Ns}} [\delta(f) + e^{-2Ns} \, \delta(1 - f)] \quad (49)$$

$$g_{ss}(f) = \frac{2\mu N}{1 + e^{-2Ns}} \frac{e^{-2Nsf}}{f(1 - f)}, \; f(1 - f) \gg \frac{1}{N}, N\mu \ll \frac{1}{\ln(1/\mu)} \quad (50)$$

$$p_{pol} \simeq 2\mu N \ln [\min(N,1/s)] \quad (51)$$

where $p_{pol} \simeq 1$ is the total probability of having a polymorphic population.

At small μN, both forms of probability density (equations 48 and 49), have singularities at $f = 0$ and $f = 1$, although the singularities are of different kinds. The two forms happen to be, to some extent, inter-changeable in the entire interval $N \ll 1/\mu$. For example, equations 48 and 49 can be shown to have, within relative error of $\sim\mu N$, the same lower momenta of the density function: the expectation value, variance, etc. The form of equation 49, although less compact, is generally more convenient to use at $N \ll 1/\mu$. The form of equation 48, on the

other hand, applies at $\mu N > 1$ as well and is suitable for studies of transition to the deterministic limit.

Neutral case: $s \ll \mu$. We start from a simple case, $s \ll \mu$, when mutations can be considered neutral. We will use the form of equation 48 for the probability density, since we want to describe both small and large populations. Setting $s = 0$ in equation 48 and normalizing the resulting expression, we get (80)

$$\rho_{ss}(f) = \frac{\Gamma(4\mu N)}{\Gamma^2(2\mu N)}[f(1-f)]^{-1+2\mu N}, s \ll \mu \qquad (52)$$

where $\Gamma(x)$ is the Euler gamma function, $\Gamma(x) = \int_0^\infty dt\, t^{x-1}e^{-t}$, and we used the identity (1)

$$\int_0^1 df\, f^{x-1}(1-f)^{y-1} = \frac{\Gamma(x)\Gamma(y)}{\Gamma(x+y)} \qquad (53)$$

The change in shape of the probability density with N is shown in Fig. 4.

The expectation values and variances of mutant frequency, f, and intrapatient distance, T, which can be obtained from equation 52 and the definitions in equations 36 to 38 are as follows:

$$\bar{f} = \frac{1}{2},\ V_f = \frac{1}{4(1+4\mu N)},$$

$$\overline{T} = \frac{2\mu N}{1+4\mu N},\ V_T = \frac{2\mu N}{(1+4\mu N)^2(3+4\mu N)}, s \ll \mu \qquad (54)$$

To evaluate the integrals over f in equations 36 and 37, we used the identities in equation 53 and $\Gamma(x+1) = x\Gamma(x)$ (1). For small populations ($\mu N \ll 1$), equations 54 yield well-known results of the neutral theory:

$$\bar{f} = 1/2,\ V_f = 1/4 \qquad (55)$$

$$\overline{T} = 2\mu N,\ V_T = 2\mu N/3,\ \mu N \ll 1,\ sN \ll 1 \qquad (56)$$

As expected, the relative standard deviation of the mutant frequency, $V^{1/2}/\bar{f}$, is on the order of 1 at $\mu N \lesssim 1$ and small in the deterministic limit, $\mu N \gg 1$ (Fig. 6b).

Case with selection: $\mu \ll s \ll 1$. As in the neutral case considered above, the probability density $\rho(f)$ shrinks as N increases. However, selection, accounted for by the factor $\exp(-2Nsf)$ in the right-hand side of equations 48 and 50, causes asymmetry of $\rho(f)$. Another important difference is the appearance of an additional asymptotic interval in N (selection-drift regime in Table 1).

For the smallest populations, $N \ll 1/s$ (drift regime), we neglect s in equation 48 and arrive at the results obtained above for the neutral case. In the opposite limit, $N \gg 1/\mu$ (the selection regime in Table 1), the probability density (equation 48) has a sharp maximum near $f = \mu/s$. Expanding $\ln \rho_{ss}$ near its maximum, one obtains a Gaussian curve (29)

$$\rho_{ss}(f) \simeq Ce^{-\frac{Ns^2}{\mu}\left(f-\frac{\mu}{s}\right)^2},\ \mu N \gg 1, s \gg \mu \qquad (57)$$

The maximum position, μ/s, is the deterministic steady-state value of the mutant frequency (see below). Expectation values and variances of f and T (equations 36 to 38) are given by

$$\bar{f} = \frac{\mu}{s},\ V_f = \frac{\mu}{2Ns^2},\ \overline{T} = 2\frac{\mu}{s},\ V_T = \frac{2\mu}{Ns^2},\ \mu N \gg 1, s \gg \mu \qquad (58)$$

In the intermediate selection-drift regime ($1/s \ll N \ll 1/\mu$), the probability density is not narrow, since equations 58 yield $V_f \gg \bar{f}$. At the same time, one is not allowed to neglect selection by putting $s = 0$ in equation 48 or equations 49 and 50. In this interval, we will analyze $\rho_{ss}(f)$ using the form given by equations 49 and 50, which can be shown to give asymptotically correct lower momenta even at $1/\ln(1/\mu) \ll N\mu \ll 1$. The function $\rho_{ss}(f)$ has three components, as shown in Fig. 5: a large peak at $f = 0$, a tiny peak at $f = 1$, and a continuous exponential tail at $f \sim 1/Ns$, which describes the density of polymorphic states. The total probability of polymorphism (equation 51) is given by a small value, $p_{pol} \simeq 2\mu N \ln(1/s)$. To obtain momenta for f and T, we substitute equation 49 into equations 36 and 37 and note that integrals of $g_{ss}(f)$ are mostly contributed by small $f \sim 1/Ns$. As a result, we obtain

$$\bar{f} = \frac{\mu}{s} + e^{-2Ns},\ V_f = \frac{\mu}{2Ns^2} + e^{-2Ns},$$

$$\overline{T} = \frac{2\mu}{s},\ V_T = \frac{2\mu}{Ns^2},\ 1/s \ll N \ll 1/\mu \qquad (59)$$

At $N \sim 1/s$, the above four values match, to an order of magnitude, the corresponding neutral values in equations 55 and 56. At $N \sim (1/s)\ln(s/\mu)$, they asymptotically match equations 58 derived above in the deterministic limit, $N \gg 1/\mu$. Remarkably, in most of the selection-drift interval, $(1/s)\ln(s/\mu) \ll N \ll 1/\mu$, the average and variance of f and T, happen to coincide with their respective deterministic formulas, even though the relative standard deviations, $V_f \mid \bar{f}$ and $V_T \mid \overline{T}$ are large, as they should be, given the shape of $\rho(f)$ (Fig. 6b).

Deterministic Dynamics and Its Boundaries

As we have shown above, the steady state is asymptotically deterministic in the limit $N \gg 1/\mu$. The purpose of this section is to consider a more general, time-dependent case. We derive the deterministic evolution equation in two independent ways, directly from deterministic first-principles, and from the stochastic equation in the limit $N \to \infty$; we solve it for an arbitrary initial condition and thus obtain the boundaries of the deterministic approximation.

Main results and discussion. In the deterministic limit, $N \to \infty$, the time-dependent probability density is given by

$$\rho(f,t) = \delta(f - f_d(t)) \qquad (60)$$

$$\frac{df_d}{dt} = M(f_d) = -sf_d(1-f_d) - \mu(2f_d - 1) \qquad (61)$$

Equation 61 represents the evolution equation for the deterministic frequency $f_d(t)$ (Fig. 7). This agrees with the meaning of $M(f)$, defined by equation 18, as the average change in f per generation (equation 20). Since the random factor, in this limit, is absent, the actual change in f naturally coincides with

the average change. The first term in the right-hand side of equation 61 describes selection for or against the minority allele and vanishes in a uniform population, $f_d = 0$ or 1. The second term in equation 61, describing mutations, does not vanish at $f_d = 0$ or 1 since mutations occur even in a uniform population. Instead, the term vanishes at $f_d = 1/2$, when the effects of forward and reverse mutations (with equal rates) cancel each other.

The solution of equation 61 in two asymptotic cases, the neutral case and the opposite case with selection, has the form

$$f(t) = \begin{cases} \frac{1}{2} + (f_o - \frac{1}{2})e^{-2\mu t}, & \text{if } \mu \ll s \\ f_{ss} + \dfrac{(f_o - f_{ss})e^{-st}}{1 + f_{ss} - f_o + (f_o - f_{ss})e^{-st}} & \text{if } \mu \gg s \end{cases} \quad (62)$$

where, in the second case, $f_{ss} = \mu/s$ (23, 24).

In the following subsection, we derive the deterministic equation 61 from the stochastic equations 1 and 2. Equation 61 can also be obtained directly from deterministic first principles. The initial set of equations appropriate for the virus population model (see "Description of the model and the evolution equation") has the form

$$\frac{dn_1}{dt} = (1 - \mu)\kappa(1 - s)n_1 + \mu\kappa n_2 - \omega n_1 \quad (63)$$

$$\frac{dn_2}{dt} = (1 - \mu)\kappa n_2 + \mu\kappa(1 - s)n_1 - \omega n_2, \quad (64)$$

where n_1 and n_2 are the numbers of mutant- and wild-type-infected cells, respectively, κ is the replication constant for the wild type, and ω^{-1} is the average life span of an infected cell. To match the virus population model where generations of infected cells change at discrete moments in steps of $\Delta t = 1$, we choose $\omega = 1$. Using the condition $n_1 + n_2 = N = const$ and the notation $f = n_1/N$, equations 63 and 64 are replaced by a single equation, equation 61.

Note that the condition $N = const$ requires that κ depends on f, as given by $\kappa - 1 = \kappa(sf - \mu)$, which explains why the resulting equation 61 is nonlinear. The reason behind this is that κ must depend on some hidden "fast" variable which adjusts quickly to relatively slow changes in f to keep N constant. An example of such a variable is the number of available target cells, which may be stable due to a balance between replenishment and killing by virus. If $f(t)$, for example, increases slowly, the average replication rate will decrease, causing, due to decreased killing, an increase in the number of target cells, which will compensate for the initial decrease in replication rate and keep N constant. There are examples of biological systems in which such a feedback mechanism is absent and the condition $N = const$ does not apply; we do not consider them in this work.

At large but finite N, the probability density has a finite width, $w(t)$, due to random drift, which is calculated in the subsection below on boundaries of deterministic approximation. The deterministic approximation remains adequate as long as the ratio $w(t)/f_d(t)[1 - f_d(t)]$ remains much less than 1. In the neutral limit, $\mu \gg s$, the deterministic boundary in N is the same as in the steady state, $N \gg 1/\mu$. In the presence of selection, $\mu \ll s$, the deterministic boundary depends, in general, on the initial condition set in the experiment. We will list results for the first three experiments in the section on experiments on evolution and observable parameters (see above), assuming that the initial value of f is known exactly, so that $w(0) = 0$:

$$w = f(1 - f) \times \begin{cases} 1/(N\mu)^{1/2}, & \text{accumulation: } f_0 = 0,\ f \ll f_{ss} \\ 1/(N\mu)^{1/2}, & \text{reversion: } f_0 = 1, f_{ss} \ll f < 1 \\ \dfrac{1}{(N_s)^{1/2}}\left[\dfrac{1 - 2f}{f(1 - f)} + 2\ln\left(\dfrac{1 - f}{f}\right)\right]^{1/2}, & \text{growth competition: } f_o = 1/2, f_{ss} \gg f < 1/2 \end{cases} \quad (65)$$

where $f \equiv f_d(t)$. We observe that in the first two experiments, when the initial population is monomorphic, the deterministic criterion is the same as in the steady state, $N \gg 1/\mu$. For the growth competition exeperiment, the criterion on N is much softer, $N \gg 1/s$, as long as f remains larger than its characteristic value in the steady state, $f \gg 1/Ns$ (cf. Fig. 5).

Deterministic dynamics. In the limit of large population numbers, $N \to \infty$, the term on the right-hand side of equation 2 corresponding to random drift is relatively small. Consequently, the density function must be narrow in f. We start by rewriting the master equation 1 and 2 in the equivalent form

$$\frac{\partial \rho}{\partial t} + M(f)\frac{\partial \rho}{\partial f} = \frac{1}{2N}\frac{\partial^2}{\partial f^2}[f(1 - f)\rho] - M'(f)\rho \quad (66)$$

where $M'(f) \equiv dM/df$ and $M(f)$ is defined by equation 18. Both terms on the right-hand side of equation 66 are relatively small: the first term is proportional to $1/N$, and the second term is much smaller than the second term on the left-hand side since $\rho(f)$ changes much more rapidly in f than $M(f)$ does. In the zero approximation in $1/N$, one can substitute 0 for the right-hand side. The resulting equation has a delta function for its partial solution, as given by equations 60 and 61. [The general solution for $\rho(f, t)$ is any linear combination of different solutions of the form of equation 60, as determined by the initial condition $\rho(f,t) \equiv \rho_0(f)$. If the initial mutant frequency is set by experiment and/or known exactly, $\rho(f,t)$ is a single delta function at all times.] In the following section, we solve equation 66 in the next approximation in $1/N$.

A solution, $f(t)$, of the deterministic equation 61 can be obtained in a general form. We rewrite equation 61 as

$$\frac{1}{s}\frac{df}{dt} = (f - f_{ss})(f - f_*) \quad (67)$$

$$f_{ss,*} = \frac{1}{2} + \frac{\mu}{s} \mp \left[\frac{1}{4} + \left(\frac{\mu}{s}\right)^2\right]^{\frac{1}{2}} \quad (68)$$

where the minus and plus signs stand for f_{ss} and f_*, respectively (here and in the rest of this section, we omit the subscript in f_d). The values of the parameters f_{ss} and f_* are restricted to the intervals $0 < f_{ss} < 1/2$ and $f_* > 1$. f_{ss} represents the mutant frequency in the steady state; f_* has no particular meaning. As one can check, f_{ss} matches its asymptotic values of 1/2 and μ/s in the respective limits $\mu \ll s$ and $\mu \ll s$ (see "Steady state"

above). Separating the variables in equation 67 and integrating, one obtains

$$f(t) = f_{ss} + \frac{(f_o - f_{ss})(f_* - f_{ss})e^{-(f_* - f_{ss})st}}{f_* - f_o + (f_o - f_{ss})e^{-(f_* - f_{ss})st}} \tag{69}$$

where $f_0 \equiv f(0)$. From equation 69, $f(\infty) = f_{ss}$. Note that the function $f(t)$ is monotonous, so that it never crosses f_{ss}. Asymptotics of equation 69 in the two cases (the neutral case and the opposite case with selection) are listed in equations 63 and 64. The characteristic time it takes to reach steady state is given, in the two respective limits in equation 62, by $1/\mu$ and $1/s$. Plots of $f(t)$ for $\mu \ll s$ for three particular initial conditions, $f_0 = 0$, 1/2, and 1, which correspond to accumulation, reversion, and growth competition experiments (see "Experiments on evolution and observable parameters" in the qualitative section of this reviews), are shown schematically in Fig. 8. The approximate expression for $f(t)$ in the accumulation experiment is especially simple,

$$f(t) = \frac{\mu}{s}(1 - e^{-st}), f_0 = 0, \mu \ll s \tag{70}$$

Boundaries of deterministic approximation. To establish conditions under which the deterministic description applies, one has to find the finite width of the probability density at finite N. For this, we use the perturbation method to solve equation 66 in the next approximation in $1/N$. We will seek a solution in the automodel form

$$\rho(f,t) = \frac{1}{w(t)} F\left(\frac{f - f_d(t)}{w(t)}\right) \tag{71}$$

where $F(u)$ is some normalized function, $\int du\, F(u) = 1$, which does not depend on time explicitly and whose width in u is on the order of 1. We assume the width of the probability density w to be much less than f_d. Later, we will obtain the interval of N in which this assumption actually holds. Since we are interested in the region of f such that $|f - f_d| \sim w$, we can replace $M(f)$ on the left-hand side of equation 66 by its linear expansion in $f - f_d$. On the right-hand side, which is already small in $1/N$, we retain only the largest terms replacing $f(1 - f) \to f_d(1 - f_d)$, $M(f) \to M(f_d)$. Substituting equation 71 into equation 66, we obtain

$$-f'(t) + M(f_d) = \left[\frac{sf_d(1 - f_d)}{2w}\right]\frac{F''(u)}{F'(u)} + \frac{1}{(Ns)^{1/2}}[w'(t) - M'(f_d)w]\frac{F(u) + uF'(u)}{F'(u)} \tag{72}$$

where f_d and w depend on time, t, and primes denote the derivatives in the corresponding arguments (shown in parentheses). We observe that the left-hand side of this expression and the bracketed terms on the right-hand side depend only on t while the factors multiplying the brackets are functions only of u. Since u and t are independent variables, equation 72 can be satisfied only if the ratio of bracketed terms is a constant (which we denote λ) and the left-hand side is identically zero. As a result, we arrive at three separate differential equations: equation 61 for $f_d(t)$ and the equation for $F(u)$ and $w(t)$:

$$\lambda F'' + uF' + F = 0 \tag{73}$$

$$\frac{dw}{dt} - M'(f_d)w - \frac{\sqrt{Ns^3}}{2\lambda}\frac{f_d(1 - f_d)}{w} = 0 \tag{74}$$

The constant, λ in the above equations can be replaced by 1, substituting $u \to \lambda^{1/2}u$ and $w \to \lambda^{-1/2}w$, which does not change the probability density (equation 71). The normalized solution of equation 73 is a Gaussian:

$$F(u) = \frac{1}{\pi^{1/2}} \exp\left[-\frac{(u - C)^2}{2}\right] \tag{75}$$

One can arbitrarily choose $C = 0$ since any other choice is equivalent to a redefinition of f_d in equation 71, which does not change the resulting probability density.

To find $w(t)$, equation 74 can be reduced to two equations with separating variables, substituting $w(t) = y(t)\phi(t)$, where y meets the equation

$$\frac{dy}{dt} - M'(f_d)y = 0 \tag{76}$$

Solving the resulting equation for $\phi(t)$ and equation 76, we obtain a general solution of the form:

$$w(t) = e^{\int M'(f)dt}\left[C + \frac{1}{N}\int dt\, f(1 - f)e^{-2\int dt\, M'(f)}\right]^{\frac{1}{2}} \tag{77}$$

where $f \equiv f_d(t)$. The integrals in t in this expression can be rewritten as integrals in f by using equation 61. As a result, the width w can be expressed in terms of the deterministic value $f(t)$ and of the initial width, $w(0)$, as given by

$$w = w(0) + \frac{|M(f)|}{N^{1/2}}\left|\int_{f_0}^{f} d\phi\, \frac{\phi(1 - \phi)}{M^3(\phi)}\right|^{1/2} \tag{78}$$

$$\text{where } (f - f_{ss})(f_o - f_{ss}) > 0. \tag{79}$$

Equation 79 is necessary for convergence of the integral in equation 78. It reflects the fact that $f(t)$ never crosses its steady-state level (see the previous subsection).

We now calculate the ratio $w/[f(1 - f)]$ for a few cases of interest. The deterministic criterion requires that this ratio be less than 1. Let us start from the steady state. Since the integral in equation 78 diverges at $f = f_{ss}$, one has to consider the upper limit of the integral, f, to be close to f_{ss}, then expand $M(f) \simeq M'(f_{ss})(f - f_{ss})$, and then evaluate the limit $f \to f_{ss}$. This yields

$$w_{ss} = f_{ss}(1 - f_{ss}) \times \begin{cases} 1/(2N\mu)^{1/2}, & \mu \ll s \\ 1/(N\mu)^{1/2}, & \mu \gg s \end{cases} \tag{80}$$

where f_{ss} is, of course, different in the two cases. The deterministic criterion is met when $N \gg 1/\mu$ (see "steady state" above). At $\mu \gg s$, as one can show from equation 78, the same condition on N applies even far from steady state. At $s \gg \mu$ and far from steady state, the condition on N depends on both f_0 and $f(t)$ (equation 65).

Stochastic Dynamics: the Drift Regime

The problems of interest are the decay of a polymorphic state and gene fixation, transition from a monomorphic state to the steady state, divergence of separated populations, and the rate of genetic turnover in the steady state.

Main results and discussion. As in the steady state, selection can be neglected if $N \ll \min(1/s, 1/\mu)$ (see "Steady state" above). In most of this interval, mutations enter only in the boundary conditions and are negligible in the polymorphic state. Dynamic experiments in this regime exhibit two main timescales: the shorter scale, at which mutation can be neglected, $t \sim N$, and a much longer scale, associated with mutations, $t \sim 1/\mu$.

Consider a polymorphic population with an initial value of f close neither to 0 nor to 1 and focus on the shorter timescale. As discussed in the qualitative section of this review, $f(t)$ drifts randomly until the population, at some point, hits a monomorphic state (Fig. 9b). In terms of the probability density, $g(f,t)$ (equation 4) spreads from the point $f = f_0$ onto the whole interval of f and then decays, as a whole, in time (Fig. 9a) as given by the following equations (32, 78):

$$g(f,t) = \left(\frac{N}{2\pi f_o(1-f_o)t}\right)^{1/2} \exp\left[-\frac{N(f-f_0)^2}{2f_0(1-f_0)t}\right], 1 \ll t \ll N \tag{81}$$

$$g(f,t) \equiv 6f_0(1-f_0)e^{-\frac{t}{N}}, t \gg N \tag{82}$$

Note the relation between distance and time following from equation 81, $f - f_0 \sim (tN)$, the same as in a gas diffusion process. At $t \gg N$, the probability is being "absorbed" by two monomorphic states, as gas is "absorbed" by two very cold walls. If, however, the initial polymorphism is very small ($f_0 \ll 1$), the manner of spread of $g(f)$ differs from the classical diffusion law:

$$g(f,t) = A f_0 \frac{2N}{t^2} e^{-\frac{2Nf}{t}}, \sqrt{(f_0 N)} \ll t \ll N \tag{83}$$

where $A \sim 1$.

Choosing $f_0 = 1/N$ and using equation 83, one can estimate the probability that a single mutant introduced into a population will ever grow to frequency f before becoming extinct and the average time of such successful growth, as given by

$$G(f) \sim \frac{1}{Nf}, t_G(f) \sim Nf \tag{84}$$

respectively. We have $G(1/N) \sim 1$, since a single allele was present to start with. The gene fixation probability is $G(1) \sim 1/N$, with the corresponding time $t_G(f) \sim N$ (34).

Transition from a monomorphic (e.g., $f = 0$) to steady state occurs in two stages, as shown in Fig. 10a. At the first stage, $t \sim N$, a thin tail of the density $g(f,t)$ spreads into the interval $0 < f < 1$ while the probability p_0 remains close to 1. (This means that some rare populations acquire an admixture of mutants.) At the second stage, $t \sim 1/\mu$, the probability $p_0(t)$ drops slowly, and $p_1(t)$ slowly increases, both approaching 1/2:

$$g(f,t) = \frac{2\mu N}{f} e^{-\frac{2Nf}{t}}, t \ll N \tag{85}$$

$$g(f,t) = \frac{\mu N}{f(1-f)}[1 + (1-2f)e^{-2\mu t}], t \gg N \tag{86}$$

$$p_{0,1} = \frac{1 \pm e^{-2\mu t}}{2} + O(\mu N) \tag{87}$$

The expectation value and variance of f, given by

$$\bar{f}(t) \simeq p_1 = \frac{1}{2}(1 - e^{-2\mu t}) \tag{88}$$

$$V_f(t) = \frac{1}{4}(1 - e^{-4\mu t}) \tag{89}$$

saturate, as they should, at the steady-state values (equation 55). Note that the time dependence of $\bar{f}(t)$ (equation 88) is the same as in the deterministic limit (equation 62). It also noteworthy that the average intrapatient distance $\overline{T}$, and its variance V_T as found from equation 86, do not depend on t at $t \gg N$. They approach their steady-state values (equation 56) much earlier than the two monomorphic states become equally probable.

The two timescales in accumulation and steady state can also be obtained from the results for the gene fixation experiment (equation 84). A mutant genome appears in the population with the small probability μN per generation. It gets fixed with the probability $G(1) \sim 1/N$. Hence, an average time between the switch from pure wild type to pure mutant and back is on the order of $N/(\mu N) = 1/\mu$. The time taken by a separate switch is much shorter and is the same as the fixation time, $t_G(1) \sim N$ (equation 84) (compare the simulation in Fig. 10a).

The longer timescale, $1/\mu$, also appears in the divergence of separated populations and the genetic turnover experiments (see "Experiments on evolution and observable parameters" in the qualitative section of this review. After two populations are isolated, at $t = 0$, from the same population, the relative genetic distance between them (equation 44) has a form

$$\overline{D}(t) = \frac{1}{2}(1 - e^{-4\mu t}) \tag{90}$$

The time correlation function for a steady state of a single population decays with time, as given by

$$K(t) = e^{-2\mu t} \tag{91}$$

Decay of the polymorphic state and gene fixation. At the smallest $N \ll 1/\mu$ and $1/s$, we use the formalism in equations 4 to 7. Neglecting selection, equation 7 becomes

$$\frac{\partial g}{\partial t} = \frac{1}{2N}\frac{\partial^2}{\partial f^2}[f(1-f)g] \tag{92}$$

which has to be solved together with boundary conditions (equations 5 and 6) and specific initial conditions, $p_0(0)$, $p_1(0)$, and $g(f, 0)$.

In this subsection, we consider an initial polymorphic population with a known mutant frequency

$$p_0(0) = p_1(0) = 0, \; g(f,0) = \delta(f - f_0) \tag{93}$$

where $f_0 \neq 0$ or 1. As shown in the next subsection, mutations (which enter the problem via equation 6) become important at much longer timescales than those involved in random drift.

Hence, we can set $\mu = 0$ in equation 6 implying that $g(f,t)$ does not diverge at the boundaries [or diverges more slowly than $1/f$ and $1/(1-f)$]. The general solution of equation 92, $g(f,t)$, can be written as a sum over its eigenfunctions $h_i(f)$ (32):

$$g(f,t) = \sum_{i=0}^{\infty} a_i\, h_i(f) e^{-\frac{\lambda_i t}{N}} \tag{94}$$

$$\frac{\partial^2}{\partial f^2}[f(1-f)h_i] = -2\lambda_i\, h_i(f) \tag{95}$$

$$a_i = \int_0^1 df\, f(1-f) h_i(f) g(f,0) \tag{96}$$

Equation 95 is equivalent to the hypergeometric equation. The eigenvalues λ_i corresponding to nondivergent solutions of equation 95 and the eigenfunctions $h_i(f)$ are given by (1)

$$\lambda_i = 1 + \frac{i(i+3)}{2},\ i = 0, 1, 2 \ldots$$

$$h_i(f) = 2\left[\frac{2i+3}{(i+1)(i+2)}\right]^{1/2} C_i^{(3/3)}(1-2f) \tag{97}$$

where $C_i^{(3/2)}(x)$ are Gegenbauer polynomials (1). The set of functions $\{h_i\}$ is orthogonal and normalized, as given by $\int_0^1 df\, f(1-f)h_i(f)h_j(f) = \delta_{ij}$. Below, we evaluate $g(f, 0)$ in asymptotic limits in time, for two cases: strong and weak initial polymorphism.

(i) Decay of strong polymorphism. Suppose that the initial population is strongly polymorphic, $f_0 \sim 1 - f_0 \sim 1$ in equation 93. For an initial period, the density $g(f,t)$ is localized in a small region of f near $f = f_0$. The factor $f(1-f)$ in equation 92 can be then approximated by a constant. It is easier to solve the resulting simplified equation directly rather than using the general solution in equation 94. Since the density is localized far from the boundaries, it is expected to have an automodel form. Substuting $g(f,t) = A(t)\mathrm{F}[\mathrm{B}(t)(f - f_0)]$ into the approximate equation 92, one arrives at equation 81. The solution applies while $t \ll N$, when the density peak remains narrow.

In the opposite limit, $t \gg N$, the sum in equation 94 can be approximated, with an exponentially small error, by its lowest term with $i = 0$. Finding λ_0 and $h_0(f)$ from equations 97 and a_0 from equations 93 and 96, we arrive at equation 82 (32, 78).

(ii) Gene fixation. We consider now a weak initial polymorphism, $f_0 \ll 1$ in equation 93. For example, $f_0 = 1/N$ corresponds to a single new genome introduced into a monomorphic population. We estimate the probability, $G(f)$, of having the new subpopulation grow to frequency f before it becomes extinct and the average time of such growth, $t_G(f)$, if this event happens. This problem has received much attention in the literature (19, 24, 34, 38, 78). A treatment based on the backward Kolmogorov equation (34) treats the final mutant frequency, f, as a constant and the initial mutant frequency, $f(0)$, as a variable. We present here a semiquantitative derivation based on the forward Kolmogorov equation. In any approach, since the range of values $f \sim 1/N$ is involved, one can estimate G only up to a numerical constant depending on the finer details of a population model (a fact not emphasized in reference 34).

At large $t \gg N$, the probability density, $g(f,t)$, still decays as given by equation 82. However, most of the decay occurs much earlier. At t such that $1 \ll t \ll N$, the density $g(f,t)$ is localized at $f \ll 1$. Unlike in the case of strong initial polymorphism, only the factor $1 - f$ in equation 92 can be replaced by a constant, 1: the factor f has to be preserved as a variable. The new automodel solution for equation 92 has the form of equation 83. Using equation 83, for the total probability of polymorphism we obtain.

$$p_{\mathrm{pol}}(t) = \int_0^{\infty} df\, g(f,t) = \frac{A}{t} \tag{98}$$

This means that the subpopulation started by a new allele at $t \sim 1$ will most probably become extinct after a few generations. The normalization coefficient A is estimated from the condition $p_{\mathrm{pol}}(t) = 1$ at $t \sim 1$, which yields $A \sim 1$. This is an estimate since the continuous approach ceases to apply at $t \sim 1$ and $f \sim 1/N$. The probability $G(f,t)$ that the frequency of new allele will exceed f at time t is given by

$$G(f,t) = \int_f^{\infty} df'\, g(f',t) \sim \frac{1}{t} e^{-\frac{2Nf}{t}} \tag{99}$$

As a function of time, the probability $G(f,t)$ has its maximum at $t = 2Nf$. The height and position of this maximum yield the desired estimates for the probability of growth $G(f)$ and the average time of growth $t_G(f)$ (equations 84).

Transition from a monomorphic to a steady state. We consider now the accumulation and reversion experiment (see "Steady state" above). Since the two alleles are symmetric when $s = 0$, it suffices to consider only one of these experiments. Let the population consist initially of the wild type only. This corresponds to initial conditions

$$p_0(0) = 1, p_1(0) = 0, g(f,0) \equiv 0 \tag{100}$$

in equation 92.

At short times t, $g(f, t)$ is localized near the left boundary, $f \ll 1$, and, as verified below, we have $p_0 \simeq 1$, $p_1 \simeq 0$. Using these facts, equations 6 and 92 can be simplified:

$$\frac{\partial g}{\partial t} = \frac{1}{2N}\frac{\partial^2}{\partial f^2}(fg),\ [fg]_{f\to 0} = 2\mu N \tag{101}$$

At the initial conditions given by equation 100, equation 101 has an automodel solution, equation 85. At $t \sim N$, the probability density spreads over the whole interval of f. The total change in p_0 can be estimated by integrating the first of equations 5 over the interval between $t \sim 1$ and $t \sim N$, which yields $1 - p_0 \sim \mu N \ln N \ll 1$. This confirms our initial guess that p_0 remains close to 1 in this interval of time. The average mutant frequency, f, and polymorphism, T (equations 36 and 38), are given by $\bar{f}(t) \simeq \mu t$, $\bar{T}(t) \simeq 2\mu t$.

At long times, $t \gg N$, the probability p_0 is slowly decaying and p_1 is slowly increasing. Since the characteristic diffusion times of $g(f,t)$ are as short as $t \sim N$ (above), the density $g(f,t)$ is at local equilibrium at any moment of time, statically following the slow changes in $p_0(t)$ and $p_1(t)$. Hence, we can put $\partial g/\partial t = 0$ in equation 92 which yields $f(1-f)\, g(f,t) = C_1(t) + C_2(t)f$.

Finding parameters $C_1(t)$ and $C_2(t)$ from equations 5 and 6, we arrive at equations 86 and 87 for $g(f)$ and $p_{0,1}$.

Divergence of separated populations and the time correlation function. Suppose that a steady-state system is split in two populations at $t = 0$, which grow quickly to the initial size. The average relative distance between the two populations, $\bar{D}$, (equation 44) is expressed via the conditional variance $V_f(t|f_0)$ for arbitrary value of f_0. In the drift regime, the task of finding $\bar{D}$ is greatly simplified because the system is almost always either purely mutant or purely wild type, so that

$$\rho_{ss}(f_0) \approx \frac{1}{2}[\delta(f_0) + \delta(1 - f_0)] \quad (102)$$

is a good approximation for $\rho_{ss}(f_0)$. (Note that this approximation cannot be used to calculate the average polymorphism $\bar{T}$: it would yield $\bar{T} = 0$. For $\bar{f}$ and V_f, however, the accuracy is sufficient.) Hence, we need to know the variance $V_f(t|f_0)$ at only two initial values, $f_0 = 0$ and $f_0 = 1$. The value of $V_f(t|0)$, was already obtained, equation 89, and from symmetry between the two alleles, we have $V_f(t|1) \equiv V_f(t|0)$. Using equations 44 and 102, we arrive at equation 90.

The time correlation function, $K(t)$, characterizing the speed of genetic turnover in a single steady-state population can be expressed in terms of the conditional expectation value $\bar{f}(t|f_0)$, as given by equation 46. Evaluation of $K(t)$ is similar to evaluation of the relative distance. The value of f_0 which mostly contributes to the right-hand side of equation 46 is $f_0 = 1$. We have $\bar{f}(t|1) = 1 - \bar{f}(t|0)$ from symmetry, where $\bar{f}(t|0)$ is already known from equation 88. Substituting $V_f^{ss} = 1/4$ from equation 55, we obtain equation 91.

Stochastic Dynamics: the Selection-Drift Regime

Main results and discussion. In this section, we consider the interval of N, i.e., $1/s \ll N \ll 1/\mu$. The accumulation experiment consists of sprouting a weak probability density tail into the interval $0 < f < 1$ from the main peak, $\delta(f)$ (Fig. 5). The relevant scales for f and t are easy to estimate from the results on gene fixation (equation 84). Consider a typical stochastic process $f(t)$, like one shown in Fig. 11. A single genome appears and grows, drifting randomly, to a frequency $f \sim 1/Ns$ (Fig. 5). Further growth is efficiently prohibited by selection. The timescale of growth is $t_G(1/Ns) = 1/s$ (equation 84). Furthermore, mutant alleles are generated in the population with frequency μN. The probability of successful growth to $f \sim 1/Ns$ is $G(f_s) \sim s \ll 1$ (equation 84). The probability of having a polymorphic state is $\sim\mu N$ (see "Steady state" above). Hence, the average time interval between such events (i.e., between high peaks in Fig. 11) is $1/\mu Ns$. The exact expressions for the average frequency, $\bar{f}$, and its variance, V_f, are

$$\bar{f}(t|0) = \frac{\mu}{s}(1 - e^{-st}) \quad (103)$$

$$V_f(t|0) = \frac{\mu}{2Ns^2}(1 - e^{-st})^2 \quad (104)$$

At $t \to \infty$, the two parameters cross over to the steady-state values obtained in equation 59. The average intrapatient distance and its variance are given by $\bar{T} \simeq 2\bar{f}$ and $V_T \simeq 4V_f$.

Note that the expectation value of the frequency $\bar{f}(t|0)$ in equation 103 coincides with the deterministic value in equation 70, although fluctuations of f are very large. The same curious result can be demonstrated for any population model in which the function $M(f)$ in Fokker-Planck equation 22 is linear in f. In the virus population model, as in many other models, the linearity condition is met asymptotically in a nearly momomorphous state, $f \ll 1$ or $1 - f \ll 1$, which is the case in the accumulation experiment. One can imagine systems, such as diploid populations with a very strong allelic dominance (34, 38), in which $M(f)$ is not linear even at $f \to 0$ and $\bar{f}(t)$ does not equal the deterministic value at small N.

The relative genetic distance between two populations split from one at $t = 0$ and the time correlation function (see below) are given by

$$\bar{D}(t) \simeq 2V_f(t|0) = \frac{\mu}{Ns^2}(1 - e^{-st})^2 \quad (105)$$

$$K(t) = e^{-st} \quad (106)$$

respectively. Note that the timescale, $1/s$, is much shorter than the timescale in the adjacent drift regime, $1/\mu$ (see "Stochastic dynamics: the drift regime" above). Crossover between these two values occurs smoothly in the interval $1/s \ll N \ll (1/s)\ln(s/\mu)$ (see "Steady state" above), as controlled by the second peak of the probability density at $f = 1$ (Fig. 5), which is exponentially small at $N \gg (1/s)\ln(s/\mu)$.

The reversion experiment (see below) exhibits a transition from the uniform mutant, $f_0 = 1$, to an almost wild-type population, $f \sim 1/Ns \ll 1$. The evolution of the probability density $\rho(f,t)$ occurs in two stages and involves two different timescales, $t \sim 1/s$ and $t \sim 1/\mu Ns$ (i.e., the same two scales as in the accumulation experiment). The first, short time is that in which a rare population becomes polymorphic. In terms of the probability density, this corresponds to a thin tail of $\rho(f,t)$ spreading into the interval $0 < f < 1$. The second, longer time is that on which a typical population switches to the wild type, i.e., the probability of the purely mutant state, $p_1(t)$, decays. At the second stage, different components of the probability density evolve, as shown in Fig. 12 and given by

$$p_1(t) = e^{-(2\mu N s)t},\ p_0(t) \simeq 1 - p_1(t),\ t \gg 1/s \quad (107)$$

$$g(f,t) = \frac{2\mu N}{f(1-f)}[p_1(t) + (1 - 2p_1(t))e^{-2Nsf}] \quad (108)$$

with a small relative error of $\sim \mu N$ (Fig. 14). The expectation values and variances of parameters f,T are given by

$$\bar{f}(t) = p_1(t)$$

$$V_f(t) = p_1(t)[1 - p_1(t)] \quad (109)$$

$$\bar{T}(t) = 4\mu N p_1(t),\ t \gg 1/s,$$

where $p_1(t)$ is given by equation 107. Note that equations 109 have a small relative error. As a result, the three parameters do not vanish at $t \to \infty$, as would follow from equations 109, but cross over to small steady state values given by equations 59 (Fig. 5). Note also that although the initial state is a pure mutant, we have a finite genetic distance, $\bar{T}(0)$, in equations 109. This is because the average polymorphism, $\bar{T}(t)$, is already saturated at $t \sim 1/s$.

The most important result is that the average waiting-for-reversion time, $1/\mu Ns$ (51), is longer than in the deterministic regime (Fig. 13). Dependence of the reversion time on N in all three intervals of N is shown schematically in Fig. 16.

Accumulation. For the sake of simplicity, we consider an interval in N, somewhat narrower in the log sense than the selection-drift interval: $(1/s)\ \ln\ (s/\mu) \ll N \ll 1/[\mu\ \ln\ (1/s]$. Then we can (i) use the more convenient formalism in equations 4 to 7 and (ii) neglect the second density peak at $f = 1$ (Fig. 5), not considering crossover to the drift regime in the interval $1/s \ll N \ll (1/s)\ \ln(s/\mu)$ (see "Steady state" above).

We seek the probability density in the form

$$p_{\text{tot}}(f,t) = p_0(t)\delta(f) + g(f,t) \tag{110}$$

where the initial state is $g(f,0) \equiv 0$ and $p_0(t) = 1$. The density $g(f,t)$ remains localized at small f and crosses over with time to the steady-state density, $g(f,\infty) = (2\mu N/f)e^{-2Nsf}$ (equation 50). This process is described by the dynamic equation and the boundary condition

$$\frac{\partial g}{\partial t} = \frac{1}{2N}\frac{\partial^2}{\partial f^2}(fg) + s\frac{\partial}{\partial f}(fg), \tag{111}$$

$$(fg)_{f\to 0} = 2\mu N \tag{112}$$

which follow from equations 6 and 7 at $f \ll 1$ and $p_0 \simeq 1$.

At short times, $t \ll 1/s$, we have $f \ll 1/(Ns)$, and the selection term in equation 111 is negligible. Hence, we can use $g(f,t)$ for the drift regime (equation 85). In principle, one could also solve equation 111 at any t, using an expansion over eigenfunctions which are related to the Laguerre polynomials (1). The lower momenta of ρ, however, can be more easily obtained without finding $\rho(f)$ explicitly. Multiplying both sides of equation 111 first by f and second by f^2 and integrating both sides over f, we get

$$\frac{d\bar{f}}{dt} = \mu - s\bar{f} \tag{113}$$

$$\frac{d\overline{f^2}}{dt} = \frac{\bar{f}}{N} - 2s\overline{f^2} \tag{114}$$

(The right-hand side of equation 111 was integrated by parts, and we used the boundary condition, equation 112.) Solving first equation 113 and then equation 114 and using the initial conditions $f(0) = V_f(0) = 0$, we arrive at equations 103 and 104 for the expectation value, $\bar{f}(t)$, and variance, $V_f(0) \simeq \overline{f^2}$.

Divergence of separated populations and the time correlation function. The time dependence of the average relative distance, $\bar{D}(t)$, between two populations isolated at $t = 0$ from a single population is given by general equation 44, in which $V_f(t|f_0)$ is defined by equations 36 and 37 with $A(f) \equiv f$ and the initial condition $\rho(f,0) = \delta(f - f_0)$. As is clear from the structure of the density function (equation 110), the main contribution to the integral in f_0 (equation 44) comes from $f_0 = 0$. Hence, using equation 104, we obtain equation 105.

The time correlation function, $K(t)$, as follows from equation 46, is contributed only from the polymorphic initial states, $f_0 \neq 0$ or 1. Substituting equations 50 and 103 and the variance, V_f^{ss}, from equation 59 into equation 46, we arrive at equation 106.

Reversion (fixation of advantageous variant). We start from equations 4 to 7 and the initial conditions $g(f, 0) \equiv 0$, $p_0(0) = 0$, and $p_1(0) = 1$. We consider only the second, most interesting stage of evolution. On this timescale, $t \sim 1/(\mu Ns)$, the probability $p_1(t)$ decays from 1 to almost 0 and $p_0(t)$ increases from 0 to almost 1. As we did with the drift regime (see above), we use the fact that the equilibration of a polymorphic state, $f \sim 1$, does not involve the mutation rate and is relatively fast. Therefore, $g(f,t)$ follows quasistatically the change in $p_0(t)$ and $p_1(t)$. Setting $\partial g/\partial t = 0$ in equation 7 and solving the resulting equation, we get

$$q(f,t) \equiv q(t)$$

$$g(f,t) = \frac{1}{f(1-f)}\left[-\frac{q(t)}{s} + A(t)e^{-2Nsf}\right] \tag{115}$$

where the coefficients $q(t)$ and $A(t)$ change slowly in time. Substituting equation 115 into the boundary conditions given by equations 5 and 6 and solving the resulting system of equations with respect to $q(t)$, $A(t)$, $p_0(t)$, and $p_1(t)$, we arrive at equations 107 and 108 and curves shown in Fig. 14.

Sampling Effects

Main results. Suppose that to obtain an experimental estimate for the intrapatient genetic distance in a population, T^*, we isolate κ sequences from the population, determine the number of nucleotide differences for each pair of sequences, and average the result over all such pairs. The difference between T^* and the actual value, T, is characterized by the standard relative error of such a measurement, ε

$$\varepsilon^2 = \frac{2}{kT} - \frac{2(2k-3)}{k(k-1)} \tag{116}$$

Equation 116 is quite general and applies to any regime or any particular experiment on genetic evolution. The value T varies randomly between populations. It is useful to know in equation 116 the representative value of T in a polymorphic population:

$$T_{\text{rep}} = \frac{1}{p_{\text{pol}}}\int_0^1 df\ 2f(1-f)g(f) \tag{117}$$

where $p_{\text{pol}} = \int_{1/N}^{1-1/N} df\ g(f)$ is the total probability of polymorphism. T_{rep} differs from the standard average $\bar{T}$ by being averaged over polymorphic states only. For instance, in the steady state, using equations 50 and 57, we get

$$T_{\text{rep}} = \begin{cases} 1/\ln N & \text{drift regime} \\ 1/[Ns\ \ln(1/s)] & \text{selection-drift regime} \\ 2\mu/s & \text{selection regime} \end{cases} \tag{118}$$

The sample size required to reach accuracy ε can be obtained by substituting these values into equation 116 (Fig. 15).

Derivations. Consider a sample of k genomes randomly selected from a population with the mutant frequency f. An experimental estimate for f, which we denote f^*, is the proportion of genomes in the sample that are mutant

$$f^* = \frac{1}{k}\sum_{i=1}^{k} v_i \tag{119}$$

where the index i numbers the genomes in the sample and the integer number v_i assumes one of two values: $v_i = 1$ if the ith genome is mutant and $v_i = 0$ if it is wild type. The probability $P(v)$ of a particular allele is given by

$$P(1) = f, \quad P(0) = 1 - f \tag{120}$$

with the expectation value $<v_i> = f$. (We will use angle brackets to denote the average over samples, reserving the overline for the average over populations.) The expectation value, $<f^*>$, as follows from equations 119 and 120, is equal to f, which confirms that f^* is a correct estimate of f. The sampling variance of the estimate, U_f, is given by

$$U_f \equiv <(f^* - f)^2> = \frac{1}{k^2}\sum_{i=1}^{k}\sum_{j=1}^{k}\langle (v_i - f)(v_j - f)\rangle \tag{121}$$

Since v_i and v_j at $i \neq j$ are independent random numbers, any term in equation 121 with $i \neq j$ reduces to a product of two averages and vanishes. Any term in equation 121 with $i = j$, as one can obtain from equation 120, is equal to $f(1 - f)$, which yields

$$U_f = f(1 - f)/k \tag{122}$$

The estimate for the frequency of polymorphic pairs, $T = 2f(1 - f)$, could be calculated, in principle, from the estimate of the mutant frequency, f^*. However, it is usually more convenient to measure T directly as a fraction of polymorphic pairs among all possible pairs of genomes from the sample. Analogously to f^*, the estimate T^* can be written in terms of numbers v_i

$$T^* = \frac{2}{k(k-1)}\sum_{i<j}[v_i(1 - v_j) + (1 - v_i)v_j] \tag{123}$$

where $k(k - 1)/2$ is the total of all possible pairs. Each polymorphic pair, $(v_i, v_j) = (0,1)$ or $(1,0)$, contributes 1 to the sum. The expectation value of T^* is $<T^* = 2f(1 - f) = T$. The sampling variance, U_T, can be obtained from equation 123 by the same method we used to obtain equation 122:

$$U_T = \langle (T^* - T)^2\rangle = \frac{2T}{k}\left(1 - T\frac{2k-3}{k-1}\right) \tag{124}$$

Since T cannot be larger than 1/2, we have $U_T > 0$ at any k. The criterion of a sufficiently large sample is that the relative error of measurement, $\varepsilon = U_T/T$, is small. From equation 124, we arrive at equation 116 for ε.

ACKNOWLEDGMENTS

We thank Joe Felsenstein for 117 useful comments.

This work was partially supported by NIH grant 1K25AI01811 and grant R35CA44385.

REFERENCES

1. **Abramowitz, M., and I. Stegun (ed.).** 1964. Handbook of mathematical functions. National Bureau of Standards, New York, N.Y.
2. **Balfe, P., P. Simmonds, C. A. Ludlam, J. O. Bishop, and A. J. Leigh Brown.** 1990. Concurrent evolution of human immunodeficiency virus type 1 in patients infected from the same source: rate of sequence change and low frequency of inactivating mutations. J. Virol. **64:**6221–6233.
3. **Bonhoeffer, S., E. C. Holmes, and M. A. Nowak.** 1995. Causes of HIV diversity. Nature **376:**125.
4. **Burns, D. P., and R. C. Desrosiers.** 1994. Envelope sequence variation, neutralizing antibodies, and primate lentivirus persistence. Curr. Top. Microbiol. Immunol. **188:**185–219.
5. **Burns, D. P. W., and R. C. Desrosiers.** 1991. Selection of genetic variants of simian immunodeficiency virus in persistently infected rhesus monkeys. J. Virol. **65:**1843–1854.
6. **Chao, L.** 1990. Fitness of RNA virus decreased by Muller's ratchet. Nature **348:**454–455.
7. **Charlesworth, B., M. Nordborg, and D. Charlesworth.** 1997. The effects of local selection, balanced polymorphism and background selection on equilibrium patterns of genetic diversity in subdivided populations. Genet. Res. **70:**155–174.
8. **Charlesworth, D., and B. Charlesworth.** 1998. Sequence variation: looking for effects of genetic linkage. Curr. Biol. **8:**R658–R661.
9. **Chavda, S. C., P. Griffin, H. L. Zhen, B. Keys, M. A. Vekony, and A. J. Cann.** 1994. Molecular determinants of the V3 loop of human immunodeficiency virus type 1 glycoprotein gp120 responsible for controlling cell tropism. J. Gen. Virol. **75:**3249–3253.
10. **Cleland, A., H. G. Watson, P. Robertson, C. A. Ludlam, and A. J. L. Brown.** 1996. Evolution of zidovudine resistance-associated genotypes in human immunodeficiency virus type 1-infected patients. J. Acquired Immune Defic. Syndr. Hum. Retrovirol. **12:**6–18.
11. **Condra, J. H., W. A. Schleiff, O. M. Blahy, L. J. Gabryleski, D. J. Graham, J. C. Quintero, A. Rhodes, H. L. Robbins, E. Roth, M. Shivaprakash, D. Titus, T. Yang, H. Teppler, K. E. Squires, P. J. Deutsch, and E. A. Emini.** 1995. In vivo emergence of HIV-1 variants resistant to multiple protease inhibitors. Nature **374:**569–571.
12. **Delwart, E. L., H. W. Sheppard, B. D. Walker, J. Goudsmit, and J. I. Mullins.** 1994. Human immunodeficiency virus type 1 evolution in vivo tracked by DNA heteroduplex mobility assays. J. Virol. **68:**6672–6683.
13. **Donis, R. O.** 1991. Muller's ratchet and flu virus. Science **253:**308–309.
14. **Duarte, E., D. Clarke, A. Moya, E. Domingo, and J. Holland.** 1992. Rapid fitness losses in mammalian RNA virus clones due to Muller's ratchet. Proc. Natl. Acad. Sci. USA **89:**6015–6019.
15. **Eigen, M., and C. K. Biebricher.** 1988. Sequence space and quasispecies distribution, p. 3–22 *In* E. Domingo, J. J. Holland, and P. Ahlquist (ed.), RNA genetics, vol. III. CRC Press, Inc., Boca Raton, Fla.
16. **Escarmis, C., M. Davila, and E. Domingo.** 1999. Multiple molecular pathways for fitness recovery of an RNA virus debilitated by operation of Muller's ratchet. J. Mol. Biol. **285:**495–505.
17. **Felsenstein, J.** 1974. The evolutionary advantage of recombination. Genetics **78:**737–756.
18. **Fisher, R. A.** 1922. On the dominance ratio. Proc. R. Soc. Edinb. **42:**321–341.
19. **Fisher, R. A.** 1930. The distribution of gene ratios for rare mutations. Proc. R. Soc. Edinb. **50:**204–219.
20. **Groenink, M., A. C. Andeweg, R. A. M. Fouchier, S. Broersen, R. C. M. van der Jagt, H. Schuitemaker, R. E. Y. de Goede, M. L. Bosch, H. G. Huisman, and M. Tersmette.** 1992. Phenotype-associated *env* gene variation among eight related human immunodeficiency virus type 1 clones: evidence for in vivo recombination and determinants of cytotropism outside the V3 domain. J. Virol. **66:**6175–6180.
21. **Haase, A. T.** 1999. Population biology of HIV-1 infection: viral and CD4+ T cell demographics and dynamics in lymphatic tissues. Annu. Rev. Immunol. **17:**625–56.
22. **Haase, A. T., K. Henry, M. Zupancic, G. Sedgewick, R. A. Faust, H. Melroe, W. Cavert, K. Gebhard, K. Staskus, Z.-Q. Zhang, P. J. Dailey, H. H. J. Balfour, A. Erice, and A. S. Perelson.** 1996. Quantitative image analysis of HIV-1 infection in lymphoid tissue. Science **274:**985–989.
23. **Haldane, J. B. S.** 1924. A mathematical theory of natural and artificial selection. Trans. Camb. Philos. Soc. **23:**19–41.
24. **Haldane, J. B. S.** 1927. A mathematical theory of natural and artificial selection. V. Selection and mutation. Proc. Camb. Philos. Soc. **23:**838–844.
25. **Ho, D. D., A. U. Neumann, A. S. Perelson, W. Chen, J. M. Leonard, and M. Markowitz.** 1995. Rapid turnover of plasma virions and CD4 lymphocytes in HIV infection. Nature **373:**123–126.
26. **Holland, J. J., J. C. De La Torre, and D. A. Steinhauer.** 1992. RNA virus populations as quasispecies. Curr. Top. Microbiol. Immunol. **176:**1–20.
27. **Holmes, E. C., L. Q. Zhang, P. Simmonds, C. A. Ludlam, and A. J. L. Leigh Brown.** 1992. Convergent and divergent sequence evolution in the surface envelope glycoprotein of human immunodeficiency virus type 1 within a single infected patient. Proc. Natl. Acad. Sci. USA **89:**4835–4839.
28. **Hudson, R. R., D. B. Boos, and N. L. Kaplan.** 1992. A statistical test for detecting geographic subdivision. Mol. Biol. Evol. **9:**138–151.
29. **Karlin, S. and J. McGregor.** 1964. On some stochastic models in genetics, p. 245–271. *In* J. Gurland (ed.), Stochastic models in medicine and biology. University of Wisconsin Press, Madison.
30. **Keys, B., J. Karis, B. Fadeel, A. Valentin, G. Norkrans, L. Hagberg, and F. Chiodi.** 1993. V3 sequences of paired HIV-1 isolates from blood and cerebrospinal fluid cluster according to host and show variation related to m clinical stage of disease. Virology **196:**475–483.
31. **Kimura, M.** 1954. Process leading to quasi-fixation of genes in natural populations due to random fluctuations of selection intensities. Genetics **39:**280–295.
32. **Kimura, M.** 1955. Solution of a process of random genetic drift with a continuous model. Proc. Natl. Acad. Sci. USA **41:**144–150.
33. **Kimura, M.** 1955. Stochastic processes and distribution of gene frequencies under natural selection. Cold Spring Harbor Symp. Quant. Biol. **20:**33–53.
34. **Kimura, M.** 1962. On the probability of fixation of mutant genes in a population. Genetics **47:**713–719.
35. **Kimura, M.** 1964. Diffusion models in population genetics. J. Appl. Probab. **1:**177–232.

36. **Kimura, M.** 1989. The neutral theory of molecular evolution and the world view of the neutralists. Genome **31:**24–31.
37. **Kimura, M.** 1994. Population genetics, molecular evolution, and the neutral theory. Selected papers. The University of Chicago Press, Chicago, Ill.
38. **Kimura, M., and T. Ohta.** 1969. The average number of generations until fixation of a mutant gene in a finite population. Genetics **61:**763–771.
39. **Kingman, J. F. C.** 1982. The coalescent. Stochastic Processes Appli. **13:**235–248.
40. **Kingman, J. F. C.** 1982. On the genealogy of large populations. J. Appl. Probab. **19A:**27–43.
41. **Kolmogorov, A.** 1931. Ueber die analitischen Methoden in der Wahrscheinlichkeitrechnung. Math. Ann. **104:**415–458.
42. **Krone, S. M., and C. Neuhauser.** 1997. Ancestral processes with selection. Theor. Popul. Biol. **51:**210–237.
43. **Lamers, S. L., J. W. Sleasman, J. X. She, K. A. Barrie, S. M. Pomeroy, D. J. Barrett, and M. M. Goodenow.** 1993. Independent variation and positive selection in env V1 and V2 domains within maternal-infant strains of human immunodeficiency virus type 1 in vivo. J. Virol. **67:**3951–3960.
44. **Lech, W. J., G. Wang, Y. L. Yang, Y. Chee, K. Dorman, D. McCrae, L. C. Lazzeroni, J. W. Erickson, J. S. Sinsheimer, and A. H. Kaplan.** 1996. In vivo sequence diversity of the protease of human immunodeficiency virus type 1: presence of protease inhibitor-resistant variants in untreated subjects. J. Virol. **70:**2038–2043.
45. **Leigh-Brown, A. J.** 1997. Analysis of HIV-1 env gene sequences reveals evidence for a low effective number in the viral population. Proc. Natl. Acad. Sci. USA **94:**1862–1865.
46. **Liu, S. L., T. Schacker, L. Musey, D. Shriner, M. J. McElrath, L. Corey, and J. I. Mullins.** 1997. Divergent patterns of progression to AIDS after infection from the same source: human immunodeficiency virus type 1 evolution and antiviral responses. J. Virol. **71:**4284–4295.
47. **Lopez-Galindez, C., J. M. Rojas, R. Najera, D. D. Richman, and M. Perucho.** 1991. Characterization of genetic variation and 3′-azido-3′-deoxythymidine-resistance mutations of human immunodeficiency virus by the RNase A mismatch cleavage method. Proc. Natl. Acad. Sci. USA **88:**4280–4284.
48. **Lukashov, V. V., C. L. Kuiken, and J. Goudsmit.** 1995. Intrahost human immunodeficiency virus type 1 evolution is related to length of the immunocompetent period. J. Virol. **69:**6911–6916.
49. **Mansky, L. M., and H. M. Temin.** 1995. Lower in vivo mutation rate of human immunodeficiency virus type 1 than that predicted from the fidelity of purified reverse transcriptase. J. Virol. **69:**5087–5094.
50. **Mayers, D. L., F. E. McCutchan, E. E. Sandersbuell, L. I. Merritt, S. Dilworth, A. K. Fowler, C. A. Marks, N. M. Ruiz, D. D. Richman, C. R. Roberts, and D. S. Burke.** 1992. Characterization of HIV isolates arising after prolonged zidovu-dine therapy. J. Acquired Immune Defic. Syndr. **5:**749–759.
51. **Maynard Smith, J. M.** 1971. What use is sex? J. Theor. Biol. **30:**319–335.
52. **Moran, P. A. P.** 1958. A general theory of the distribution of gene frequencies. I. Overlapping generations. II. Non-overlapping generations. Proc. R. Soc. London Ser. **B 149:**102–116.
53. **Muller, H. J.** 1932. Some genetic aspects of sex. Am. Nat. **66:**118–128.
54. **Nei, M.** 1972. Genetic distance between populations. Am. Nat. **106:**283–292.
55. **Neuhauser, C., and S. M. Krone.** 1997. The genealogy of samples in models with selection. Genetics **145:**519–534.
56. **Nietfield, W., M. Bauer, M. Fevrier, R. Maier, B. Holzwarth, R. Frank, B. Maier, Y. Riviere, and A. Meyerhans.** 1995. Sequence constraints and recognition by CTL of an HLA-B27-restricted HIV-1 gag epitope. J. Immunol. **154:**2189–2197.
57. **Nordborg, M., B. Charlesworth, and D. Charlesworth.** 1996. The effect of recombination on background selection. Genet. Res. **67:**159–174.
58. **Novella, I. S., M. Cilnis, S. F. Elena, J. Kohn, A. Moya, E. Domingo, and J. J. Holland.** 1996. Large-population passages of vesicular stomatitis virus in interferon-treated cells select variants of only limited resistance. J. Virol. **70:**6414–6417.
59. **Novella, I. S., D. K. Clarke, J. Quer, E. A. Duarte, C. H. Lee, S. C. Weaver, S. F. Elena, A. Moya, E. Domingo, and J. J. Holland.** 1995. Extreme fitness differences in mammalian and insect hosts after continuous replication of vesicular stomatitis virus in sandfly cells. J. Virol. **69:**6805–6809.
60. **Novella, I. S., S. F. Elena, A. Moya, E. Domingo, and J. J. Holland.** 1995. Size of genetic bottlenecks leading to virus fitness loss is determined by mean initial population fitness. J. Virol. **69:**2869–2872.
61. **Ochman, H., and A. C. Wilson.** 1987. Evolution in bacteria: evidence for a universal substitution rate in cellular genomes. J. Mol. Evol. **26:**74–86. (Erratum, **26:**377.)
62. **Perelson, A. S., A. U. Neumann, M. Markowitz, J. M. Leonard, and D. D. Ho.** 1996. HIV-1 dynamics in vivo: virion clearance rate, infected cell life span, and viral generation time. Science **271:**1582–1586.
63. **Raikh, M. E., and I. M. Ruzin.** 1991. Transmittance fluctuations in randomly non-uniform barriers and incoherent mesoscopics, p. 301–354. *In* B. L. Altshuler, P. A. Lee, and R. A. Webb (ed.), Quantum phenomena in mesoscopic systems. North Holland, Amsterdam, The Netherlands.
64. **Reinhart, T. A., M. J. Rogan, A. M. Amedee, M. Murphey-Corb, D. M. Rausch, L. E. Eiden, and A. T. Haase.** 1998. Tracking members of the simian immunodeficiency virus deltaB670 quasispecies population in vivo at single-cell resolution. J. Virol. **72:**113–120.
65. **Rodrigo, A. G. and J. Felsenstein.** 1999. Coalescent approaches to HIV population genetics, p. 233–272. *In* K. Crandall (ed.), Molecular evolution of HIV. John Hopkins University Press, Baltimore, Md.
66. **Rodrigo, A. G., E. G. Shpaer, E. L. Delwart, A. K. Iversen, M. V. Gallo, J. Brojatsch, M. S. Hirsch, B. D. Walker, and J. I. Mullins.** 1999. Coalescent estimates of HIV-1 generation time in vivo. Proc. Natl. Acad. Sci. USA **96:**2187–2191.
67. **Rouzine, I. M., and J. M. Coffin.** 1999. Linkage disequilibrium test implies a large effective population number for HIV in vivo. Proc. Natl. Acad. Sci. USA **96:**10758–10763.
68. **Rouzine, I. M., and J. M. Coffin.** 1999. Search for the mechanism of genetic variation in the *pro* gene of human immunodeficiency virus. J. Virol. **73:** 8167–8178.
69. **Salminen, M. O., C. Koch, E. Sanders-Buell, P. K. Ehrenberg, N. L. Michael, J. K. Carr, D. S. Burke, and F. E. McCutchan.** 1995. Recovery of virtually full-length HIV-1 provirus of diverse subtypes from primary virus cultures using the polymerase chain reaction. Virology **213:**80–86.
70. **Seibert, S. A., C. Y. Howell, M. K. Hughes, and A. L. Hughes.** 1995. Natural selection on the gag, pol, and env genes of human immunodeficiency virus 1 (HIV-1). Mol. Biol. Evol. **12:**803–813.
71. **Shankarappa, R., J. B. Margolick, S. J. Gange, A. G. Rodrigo, D. Upchurch, H. Farzadegan, P. Gupta, C. R. Rinaldo, G. H. Learn, X. He, X. L. Huang, and J. I. Mullins.** 1999. Consistent viral evolutionary changes associated with the progression of human immunodeficiency virus type 1 infection. J. Virol. **73:**10489–10502.
72. **Shklovski, B. I., and A. L. Efros.** 1984. Electronic properties of doped semiconductors. Springer-Verlag KG, Berlin, Germany.
73. **Takahashi, H., R. Houghten, S. D. Putney, D. H. Margulies, B. Moss, R. N. Germain, and J. A. Berzofsky.** 1989. Structural requirements for class I MHC molecule-mediated antigen presentation and cytotoxic T cell recognition of an immunodominant determinant of the human immunodeficiency virus envelope protein. J. Exp. Med. **170:**2023–2035.
74. **Wain-Hobson, S., and M. Sala.** 1999. Drift and conservatism in RNA virus evolution: are they adapting or merely changing? p. 115–140. *In* E. Domingo, R. Webster, and J. Holland (ed.), Origin and evolution of viruses. Academic Press Ltd., London, United Kingdom.
75. **Watterson, G. A.** 1962. Some theoretical aspects of diffusion theory in population genetics. Ann. Math. Stat. **33:**939–957.
76. **Watterson, G. A.** 1975. On the number of segregating sites in genetical models without recombination. Theor. Popul. Biol. **7:**256–276.
77. **Wei, X., S. Ghosh, M. E. Taylor, V. A. Johnson, E. A. Emini, P. Deutsch, J. D. Lifson, S. Bonhoeffer, M. A. Nowak, B. H. Hahn, M. S. Saag, and G. M. Shaw.** 1995. Viral dynamics in human immunodeficiency virus type 1 infection. Nature **373:**117–122.
78. **Wolfs, T. F. W., J.-J. de Jong, H. van den Berg, J. M. G. H. Tijnagel, W. J. A. Drone, and J. Goudsmit.** 1990. Evolution of sequences encoding the principal neutralization epitope of human immunodeficiency virus 1 is host dependent, rapid, and continuous. Proc. Natl. Acad. Sci. USA **87:**9938–9942.
79. **Wolfs, T. F. W., G. Zwart, M. Bakker, M. Valk, C. L. Kuiken, and J. Gousmit.** 1991. Naturally occurring mutations within HIV-1 V3 genomic RNA leads to antigenic variation dependent on a single amino acid substitution. Virology **185:**195–205.
80. **Wright, S.** 1931. Evolution in Mendelian populations. Genetics **16:**97–159.
81. **Wright, S.** 1945. The differential equation of the distribution of gene frequencies. Proc. Natl. Acad. Sci. USA **31:**382–389.
82. **Wright, S.** 1945. Tempo and mode in evolution: a critical review. Ecology **26:**415–419.
83. **Yamaguchi, Y., and T. Gojobori.** 1997. Evolutionary mechanisms and population dynamics of the third variable envelope region of HIV within single hosts. Proc. Natl. Acad. Sci. USA **94:**1264–1269.
84. **Yoshimura, F. K., K. Diem, G. H. Learn, Jr., S. Riddell, and L. Corey.** 1996. Intrapatient sequence variation of the *gag* gene of human immunodeficiency virus type 1 plasma virions. J. Virol. **70:**8879–8887.
85. **Zhang, L., P. J. Dailey, T. He, A. Gettie, S. Bonhoeffer, A. S. Perelson, and D. D. Ho.** 1999. Rapid clearance of simian immunodeficiency virus particles from plasma of rhesus macaques. J. Virol. **73:**855–860.
86. **Zhang, L. Q., P. MacKenzie, A. Cleland, E. C. Holmes, A. J. Leigh Brown, and P. Simmonds.** 1993. Selection for specific sequences in the external envelope protein of human immunodeficiency virus type 1 upon primary infection. J. Virol. **67:**3345–3356.