Journal of Statistical Physics, Vol. 49, Nos. 5/6, 1987

A Model of Natural Selection That Exhibits a Dynamic Phase Transition

Ed Weinberger[1]

Received January 13, 1987; revision received July 30, 1987

A simple, stochastic model is developed of an asexual biological population that is undergoing natural selection. It is then observed that the size of the population, like the temperature parameter in the simulated annealing algorithm, is a measure of the amount of randomness to be allowed in the system. Exploiting the formal analogy between the two processes, it is shown that the distribution of different types of organisms in the population model converges to a stationary distribution if the population is growing more slowly than $O(\ln t)$ ("annealing"), but can fail to converge at all if the population is growing faster than $O(\ln t)$ ("quenching"). The results may be related to the "historical accidents" that permeate biological structures.

KEY WORDS: Evolutionary theory; natural selection; simulated annealing; dynamic phase transitions.

1. INTRODUCTION

The theory of natural selection that forms the basis for the modern theory of biological evolution consists of two distinct subprocesses: random variation in biological fitness and differential reproductive success based on this variation.[(1)] Nonspecialists in evolutionary theory often have difficulty with these ideas because the simple physical systems that are more familiar to them tend to lose, rather than gain, complexity as time progresses. They argue that, if mutation is truly a random process, then harmful mutations must occur. One would think that random perturbations of a complex biological system would almost certainly interfere with its functioning in

[1] Department of Biochemistry and Biophysics, University of Pennsylvania, Philadelphia, Pennsylvania 19104.

much the same way as random replacement of instructions would destroy the coherence of most computer programs. Certainly, many of the human mutations that produce qualitative changes lead to either "birth defects," "hereditary diseases," or "cancer." On the other hand, true improvements in fitness—superior strength, intelligence, or acuity, for example—seem to be quantitative changes that may not represent mutations at all. Another apparent problem with natural selection is that truly useful evolutionary adaptations, such as the development of wings from arms, must have taken many generations. If these changes are to be explained by natural selection, then each of the transitional organisms must be biologically fitter than their competitors. In other words, even half a wing must give the organism a selective advantage before the full wing can develop.

One way of restating many of the objections given above is to note that the "space" of possible biological organisms almost certainly contains many local fitness maxima. It is well known that finding the global maximum in problems with many local maxima is difficult, especially with the local search strategy that is implicit in natural selection. It is therefore hard to see how evolution can produce the best of all possible organisms. On the other hand, it is hard to argue that evolution does not exhibit some optimum-seeking behavior. In particular, there is an intriguing resemblance between simulated annealing, a Monte Carlo optimization procedure, and natural selection. There are probably many natural selection models that share features with annealing (see, for example, Ref. 2 for a discussion of the relationship between annealing and a deterministic model of natural selection due to Eigen[3]). For each, the fundamental similarity is that they are nonstationary processes that are regulated by order parameters that are themselves altered by the dynamics of the process. It is the purpose of this paper to explore this similarity by means of a particular selection model to be discussed below. Using methods that are useful in proving global convergence of the simulated annealing algorithm, I show that the relative frequencies of the various species in a model system to be described below can fail to achieve a stationary distribution if the population is growing faster than logarithmically in time, but that a stationary distribution is achieved if the population is growing more slowly in time. This is the dynamic phase transition promised in the title of this paper. I then describe conditions under which the stationary distribution of the relative frequencies does indeed have unit "probability mass" associated with the globally fittest species. The presentation also includes some thoughts about what these results might have to do with real biological systems.

2. THE SELECTION MODEL. A FINITE STATE, NONSTATIONARY MARKOV CHAIN

I start by describing the model of natural selection to be used in the subsequent discussions. Although the model might seem simplistic, it is in the spirit of other work in the field, most notably that of Gillespie.[4] I choose to consider the present model, rather than the better known Wright–Fisher diffusion model, for two reasons: First, I wish to consider the population dynamics of a large number of species under the influence of both mutation and selection, for which the Wright–Fisher theory is mathematically intractable. Second, the Wright–Fisher theory is strictly justifiable only if the selective advantage is on the order of the reciprocal of the population size. Because I plan to study the dependence of the dynamics in the limit of large population size, the Wright–Fisher theory would only apply if the selective advantages under consideration approach zero. However, the reader familiar with the Wright–Fisher theory will note that the results of that theory are recoverable from the present model upon appropriate rescaling and by taking the appropriate limits. I also note that the details of the model are important only in deriving a single formula, the probability that a species s, initially present in small numbers, eventually takes over the population. I present an argument that this formula agrees qualitatively with that derived from the more detailed and generally accepted Wright–Fisher diffusion model.

The present consists of a large, but finite number of different types of asexually replicating entities (i.e., different asexual "species"). With each species s we associate a specific value of a fitness parameter r_s. If each species were allowed to replicate without constraint, r_s would be the rate constant for the exponential growth that would occur. However, a central feature of the model (and real living systems) is that constraints are present; in particular, we assume that these constraints impose a carrying capacity of $N(t)$ organisms on the supporting ecosystem. We will assume that $N(t)$ increases slowly with time—so slowly, in fact, that N can be treated as essentially constant over the time period required for a mutation to either become fixed in the population or disappear from it. (This situation might arise, for example, when a series of mutations slowly allows an increase in a species' resource utilization.) We include the effects of mutation in the model by associating with each pair of species s and s' a small probability $\mu_{s,s'}$ that an attempted replication of an s individual produces an s' individual instead. We note that the biologically interesting case is when $\mu_{s,s'}$ is allowed to be zero for certain choices of s and s'. Because mutation is assumed to be such a rare event relative to the fixation/disappearance time for mutations, all individuals in the population will almost all of the time be of the same species. The mutation process can therefore be approximated by a sequence of N independent and identically distributed Bernoulli trials on the N individuals in the population, each with success probability $\mu_{s,s'}$.

We now are ready to derive the "acceptance function" $A(s, s')$, which is the probability that species s' takes over the population after s, given that mutations to s' occur. We assume that, initially, each of the N organisms in the population are of species s, and that, due to random mutation or migration, some small number of organisms of one other type s' are introduced into the population. We make two additional simplifying assumptions: First, we assume that the only time that mutation or migration takes place is when mutants or migrants are introduced into a previously "pure" population. (This assumption, which is perhaps less natural than that of a constant mutation rate for all time, including those times when the population is heterogeneous, deserves some discussion. One possible justification for the former assumption is that the environment is transiently mutagenic, as would be the case, for example, during unusually pronounced solar or cosmic ray activity. Alternatively, we could imagine that migration can only take place at certain rare intervals, because, for example, a land bridge, an unusually long draught, etc., make possible the crossing of normally impassible bodies of water.) We assume further that we can approximate the number of new organisms introduced, which is binomially distributed, by its expected value $\mu_{s,s'}N$. The numbers of wild-type individuals $X_s(t)$ and mutant individuals $X_{s'}(t)$, that appear in the population subsequently is assumed to be given by a birth-and-death process that conserves the total number N of organisms in the population. In some small time interval dt the organisms in state s each have probability $r_s\,dt$ of reproducing, thereby increasing their number by one. Similarly, the organisms in state s' each have probability $r_{s'}\,dt$ of reproducing. As always in such situations, it is presumed that dt is so short that at most one of these events happens in the time interval dt. In order to preserve the total number of organisms, one of the N organisms currently in the population is selected at random to die at the same instant that the new organism is born. We can then derive a system of differential equations that describes the time evolution of

$$P_n(t) = Pr\{X_s(t) = n\}$$

We observe that these probabilities change only because an organism of type s is replaced by one of type s' and vice versa. We have, for all $0 < n < N$,

$$\begin{aligned}&\Pr\{\text{type } s \text{ organism replacing}\\ &\qquad \text{one of type } s' \text{ in } dt \mid X_s(t)=n\}\\ &\qquad = \lambda_n = r_s n(1-n/N)\,dt\end{aligned}$$

and

$$\begin{aligned}&\Pr\{s' \text{ organism replacing } s \text{ organism in } dt \mid X_s(t)=n\}\\ &\quad = \mu_n\\ &\quad = r_{s'}(N-n)(n/N)\,dt\\ &\quad = r_{s'}n(1-n/N)\,dt, \qquad 0<n<N\end{aligned}$$

When $n=0$ or $n=N$, it is impossible for the population to change, so that $\lambda_0=\mu_0=\lambda_N=\mu_N=0$. With explicit expressions for λ_n and μ_n in hand for all n between 0 and N, we can substitute in the well-known evolution equations for birth-and-death processes,[5]

$$\begin{aligned}dP_0(t)/dt&= -\lambda_0 P_0(t)+\mu_1 P_1(t)\\ dP_n(t)/dt&=\lambda_{n-1}P_{n-1}(t)-(\lambda_n+\mu_n)\,P_n(t)+\mu_{n+1}P_{n+1}(t), \quad 0<n<N\\ dP_N(t)/dt&=\lambda_{N-1}P_{N-1}(t)-\mu_N P_N(t)\end{aligned}$$

to obtain

$$\begin{aligned}\frac{dP_0}{dt}&=r_{s'}\left(\frac{1-1}{N}\right)P_1(t)\\ \frac{dP_n(t)}{dt}&=r_s(n-1)\left(1-\frac{n-1}{N}\right)P_{n-1}(t)-(r_s+r_{s'})\,n\left(\frac{1-n}{N}\right)P_n(t)\\ &\quad +r_{s'}(n+1)\left(1-\frac{n+1}{N}\right)P_{n+1}(t), \qquad 0<n<N\\ \frac{dP_N(t)}{dt}&=r_s\frac{N-1}{N}P_{N-1}(t)\end{aligned}$$

Our primary interest in these equations is the computation of $\lim_{t\to\infty}P_0(t)$, the probability of eventual extinction of the wild type (or eventual takeover by the mutant). This probability turns out to be easy to compute by a consideration of the "embedded random walk." This is the random walk obtained from the birth-and-death process by forgetting about time. This random walk has transition probabilities[5]

$$\begin{aligned}p_{s,s'}&=\Pr\{\text{taking one step to the right starting at } n\}\\ &=\frac{\lambda_n}{\lambda_n+\mu_n}=\frac{r_s}{r_s+r_{s'}}\end{aligned}$$

and

$$\begin{aligned} q_{s,s'} &= 1 - p_{s,s'} \\ &= \Pr\{\text{taking one step to the left starting at } n\} \\ &= \frac{\mu_n}{\lambda_n + \mu_n} = \frac{r_{s'}}{r_s + r_{s'}} \end{aligned}$$

The probability of fixation of the mutation in the population is then given by the well-known gambler's ruin formula,[6]

$$\begin{aligned} &\Pr\{\text{fixation by mutant}\} \\ &\quad = 1 - \Pr\{\text{extinction of mutant}\} \\ &\quad = 1 - \frac{(q_{s,s'}/p_{s,s'})^N - (s_{s,s'}/p_{s,s'})^{u_{s,s'}N}}{(q_{s,s'}/p_{s,s'})^N - 1} \\ &\quad = \frac{(r_s/r_{s'})^{\mu_{s,s'}N} - 1}{(r_s/r_{s'})^N - 1} \end{aligned}$$

Note that $\mu_{s,s'}$ is typically much smaller than $1/N$ and that $r_s \approx r_{s'}$. These observations justify the two time scales that are employed in this analysis. In any case, it is this last expression that is the acceptance function $A_N(s, s')$.

When $r_s - r_{s'} = O(1/N)$, one can also model the differential growth of the mutant and wild-type populations as a diffusion process (Wright–Fisher diffusion). This model is well known in the literature of population biology; see references such as Ref. 7, from which the needed result will be quoted. If we assume that $r_s - r_{s'} = \sigma/N$ and that the initial fraction of mutants is $\mu_{s,s'}$, then

$$\Pr\{\text{fixation of mutant}\} = \frac{1 - e^{2\sigma\mu_{s,s'}}}{1 - e^{2\sigma}}$$

If $\sigma = O(1)$, then the corresponding formula from the model given above is

$$\Pr\{\text{fixation of mutant}\} = \frac{r_s/r_{s'})^{u_{s,s'}N} - 1}{(r_s/r_{s'})^N - 1} \to \frac{1 - e^{\sigma\mu_{s,s'}}}{1 - e^{\sigma}}$$

As will become clear from the subsequent discussion, the factor of 2 that distinguishes the two formulas does not qualitatively affect the conclusions that will be drawn from them.

3. THE ANALOGY WITH SIMULATED ANNEALING

I turn now to a discussion of the analogy with simulated annealing. This algorithm, first proposed by Kirkpatrick *et al.*,[8] has been used successfully on so many different applications that it can be considered a general-purpose optimization algorithm (see Ref. 9 for a partial list of applications). There is a wealth of description of the details of this algorithm in the literature.[8,9] More relevant to the present discussion are the qualitative features of the algorithm, which are reflected in the way the algorithm is named. Annealing is a well-known method of hardening metals that involves slow cooling from the liquid state. Initially, the metal atoms move about rapidly in the metal due to random thermal activity. The metal atoms encounter optimum energy configurations—minima in this case—and settle into them as the metal cools and its atoms lose energy. The atoms are more likely to be caught by an energy minimum that is "wide and deep" in phase space than by one that is "small and shallow." Therefore, the metal atoms are more likely to settle into a stable configuration (with close to maximum hardness) than a relatively unstable one.

Although a "temperature" parameter is not an explicit part of the natural selection model, something like simulated annealing can be expected to occur. Initially, a mutant organism can take over the population with relative ease, because the population contains only a small number of suboptimal organisms. The mutant organism will therefore be "accepted" relatively often, corresponding to the "high-temperature" regime in simulated annealing. As the number and sophistication of the organisms already in the population increase, a mutation will have more difficulty becoming fixed in the population. "Acceptance" of mutants therefore occurs less frequently as time passes, corresponding to the low-temperature regime in annealing.

4. CONVERGENCE PROPERTIES OF THE MODEL

One implication of this analogy is that the same methods that have been used to study the convergence properties of simulated annealing can be used to study the convergence properties of the nonstationary Markov chain defined by the natural selection model. *A priori*, it is not so obvious whether this Markov chain will, from *any* initial state, find the state s_M with the largest fitness parameter r_M. There will almost certainly be local maxima in the function that maps states to fitness parameters, because there will almost certainly be states other than s_M whose fitnessparameters are higher than all nearby states. It is possible that the population can increase sufficiently rapidly that it gets stuck in one of these local maxima. The analogous physical situation is quenching (rapid cooling), as opposed to annealing (slow cooling).

These issues can be resolved via an application of some basic ideas about the long-time behavior of nonstationary Markov chains. Strictly speaking, our model may make no sense in the long-time limit, because the population size may diverge. However, it is unlikely that the long, but finite-time behavior of the model will differ significantly from that of the long-time limit. We therefore consider the limiting behavior of the model as an approximation.

Two kinds of convergence of nonstationary Markov chains have been identified.[10] The first, known as *weak ergodicity*, requires that the process exhibit a loss of memory in the sense that, for all fixed t_0,

$$\lim_{t \to \infty} \min_{f,g} \|\mathbf{f}\mathbf{P}_{N(t_0)}\mathbf{P}_{N(t_0+1)} \cdots \mathbf{P}_{N(t)} - g\mathbf{P}_{N(t_0)}\mathbf{P}_{N(t_0+1)} \cdots \mathbf{P}_{N(t)}\| = 0$$

Here, $\mathbf{f}$ and $\mathbf{g}$ are starting probability vectors, and $\mathbf{P}_{N(t)}$ is the transition matrix for the Markov process at time t. Weak ergodicity is equivalent to the condition that $\prod_{\theta=t_0}^{t} \mathbf{P}_{N(\theta)}$ approaches a matrix with identical rows at $t \to \infty$. For finite-dimensional, stationary Markov chains, weak ergodicity and the more familiar concept known simply as ergodicity are equivalent. An alternative characterization of weak ergodicity that we need in order to prove our first convergence result requires the introduction of the *delta coefficient* $\delta(\mathbf{P})$ of the matrix $\mathbf{P}$. This quantity is defined as

$$\delta(\mathbf{P}) = \frac{1}{2} \max_{i,k} \sum_{j} |p_{ij} - p_{kj}|$$

where $\mathbf{P} = [p_{ij}]$. It is a measure of how close a nonnegative matrix is to a matrix with identical rows. It is clear that the condition

$$\lim_{t \to \infty} \delta\left(\prod_{t_0=\theta}^{t} \mathbf{P}_{N(\theta)}\right) = 0$$

for all t_0 is sufficient, but not necessary for weak ergodicity. We use, instead, a necessary and sufficient condition that is stated, not in terms of $\delta(\mathbf{P})$, but in terms of $\alpha(\mathbf{P}) \equiv 1 - \delta(\mathbf{P})$, the so-called *ergodic coefficient of Dobrushin.* The statement is then as follows[10]:

A nonstationary Markov chain is weakly ergodic iff there exist integers $0 \leqslant t_0 < t_1 < t_2 \cdots < t_n < \cdots$ such that

$$\sum_{n=0}^{\infty} \alpha\left(\prod_{t=t_n}^{t_{n+1}} \mathbf{P}_{N(t)}\right) = \infty \tag{1}$$

This theorem allows us to prove our first convergence result:

The nonstationary Markov chain $\mathscr{M}$ describing the selection model given above is weakly ergodic if $N(t) \leqslant \ln t/(\ln r_M - \ln r_m)$, where r_M and r_m are the maximum and minimum fitness parameters, respectively. However, if $N(t) > C \ln t/(\ln r_M - \ln r_m)$, for some constant C, $\mathscr{M}$ can fail to be weakly ergodic.

We prove the first part of this result by noting first that, if $N(t) = \ln t/(\ln r_M - \ln r_m)$, then

$$\begin{aligned} A_N(s, s') &= \frac{(r_s/r_{s'})^{\mu_{s,s'}N} - 1}{(r_s/r_{s'})^N - 1} \\ &= \frac{t^{(\mu_{s,s'} \ln r_s - \ln r_{s'})/(\ln r_M - \ln r_m)} - 1}{t^{(\ln r_s - \ln r_{s'})/(\ln r_M - \ln r_m)} - 1} \end{aligned}$$

which is, to leading order as $t \to \infty$,

$$= \begin{cases} t^{-(1-\mu_{s,s'})(\ln r_s - \ln r_{s'})/(\ln r_M - \ln r_m)}, & \text{if} \quad r_s > r_{s'} \\ 1 - t^{-\mu_{s,s'}(\ln r_{s'} - \ln r_s)/(\ln r_M - \ln r_m)}, & \text{if} \quad r_{s'} > r_s \end{cases}$$

Here, we ignore the unlikely possibility that $r_s = r_{s'}$. We *define* the transition probability from s to s' to be 0 when $\mu_{s,s'} = 0$ in accordance with the relationship $p^N_{s,s'} = \mu_{s,s'} A_N(s, s')$. If $\mu_{s,s'} > 0$, then

$$0 < \mu_{s,s'} \frac{\ln r_{s'} - \ln r_s}{\ln r_M - \ln r_m} < (1 - \mu_{s,s'}) \frac{\ln r_{s'} - \ln r_s}{\ln r_M - \ln r_m} < 1 \tag{2}$$

for $r_s < r_{s'}$.

We verify that (1) holds in the case where $t_n = n$, assuming, without loss of generality, that the states are ordered such that $r_0 = r_M > r_1 > r_2 > \cdots > r_N = r_m$. We can then refer to each state by its index. Using this notation, we estimate the sum

$$\begin{aligned} \sum_j |p_{ij} - p_{kj}| = &\sum_{j: r_i, r_k > r_j} |p_{ij} - p_{kj}| + |p_{ii} - p_{ki}| + |p_{ik} - p_{kk}| \\ &+ \sum_{j: r_i > r_j > r_k} |p_{ij} - p_{kj}| + \sum_{j: r_j > r_i, r_k} |p_{ij} - p_{kj}| \end{aligned} \tag{3}$$

Here, the notation j: $r_i, r_k > r_j$ means "sum over those states j with fitness parameter r_j less than $\min(r_i, r_k)$." The notations j: $r_i > r_j > r_k$ and j: $r_j > r_i, r_k$ are defined similarly. We have assumed that $i < k$, which we can do without loss of generality in computing the ergodic coefficient. Estimating each term shown above via (2), we see that, for sufficiently large t and sufficiently small μ,

$$\sum_{j:r_i,r_k>r_j} |p_{ij}-p_{kj}| = O(t^{-\alpha}) \tag{4a}$$

$$|p_{ii}-p_{ki}| = 1 - \sum_{j:r_i<r_j} \mu_{ij} - \mu_{ki} + O(t^{-b}) \tag{4b}$$

$$|p_{kk}-p_{ik}| = 1 - \sum_{j:r_k<r_j} \mu_{kj} + O(t^{-c}) \tag{4c}$$

$$\sum_{j:r_i>r_j>r_k} |p_{ij}-p_{kj}| = \sum_{j:r_i>r_j>r_k} \mu_{kj} + O(t^{-d}) \tag{4d}$$

$$\sum_{j:r_i,r_k<r_j} |p_{ij}-p_{kj}| = \sum_{j:r_i,r_k<r_j} |\mu_{ij}-\mu_{kj}| + O(t^{-e}) \tag{4e}$$

where $0<a, b, c, d, e<1$. If we observe that

$$\sum_{j:r_k<r_j} \mu_{kj} = \sum_{j:r_i>r_j>r_k} \mu_{kj} + \sum_{j:r_i<r_j} \mu_{kj}$$

and if we let $\lambda = \min(a, b, c, d, e)$, then we see that

$$\sum_j |p_{ij}-p_{kj}| = 2 - \mu_{ki} - \sum_{j:r_i<r_j} [(\mu_{ij}+\mu_{kj}) - |\mu_{ij}-\mu_{kj}|] + O(t^{-\lambda})$$

We then have the following estimate for the ergodic coefficient:

$$\begin{aligned}\alpha(\mathbf{P}_{N(t)}) &= -\frac{1}{2}\max_{i,k}\left\{-\mu_{ik} - \sum_{j:r_i<r_j} [(\mu_{ij}+\mu_{kj}) - |\mu_{ij}-\mu_{kj}|] + O(t^{-\lambda})\right\}\\ &\geqslant \frac{1}{2}\min_{i,k} \mu_{ki} + \frac{1}{2}\min_{i,k}\left\{\sum_{j:r_i<r_j} [(\mu_{ij}+\mu_{kj}) - |\mu_{ij}-\mu_{kj}|] + O(t^{-\lambda})\right\}\end{aligned}$$

The quantities in square brackets are always nonnegative if the μ's are nonnegative, so that the sum of these terms attains its minimum value of zero if we take $i=0=M$. We continue to employ the subscript M, rather than replacing it by zero, to emphasize that this subscript is associated with the maximum reproduction rate. It is clear that (1) holds if $\min_{i,k}\mu_{ik}>0$. Indeed, if the matrix $\mathbf{M}=[\mu_{ik}]$ is ergodic and l is the smallest integer such that the matrix $\mathbf{M}^l=[m_{ik}]$ is strictly positive, we could have taken $t_n=ln$, which would guarantee that $m_{ik}>0$ and the applicability of (1). It is, however, the nonergodic case that includes the possibility of local optima, so that we must consider this last case in more detail. We saw previously that each of the terms in (3) has a component of order $O(t^{-\lambda})$, where $0<\lambda<1$. We can conclude that (1) still holds if we can show that the ergodic coefficient, which is approximately the sum of these leading order terms, is $O(t^{-\varepsilon})$, where $0<\varepsilon<1$.

To make this more detailed analysis, we define k^* as that value of k such that the expression for the ergodic coefficient is minimized (The argument of the preceding paragraph implies that this definition can be made independent of i.). We then consider two subcases, depending upon whether k^* is or is not equal to 1. [We assumed, without loss of generality, that $i \leqslant k$. If $i = k$, $\delta(\mathbf{P}_{N(t)}) = 0$.] We assume first that $k^* > 1$ and reconsider (4b). We have, to leading order,

$$|p_{MM} - p_{k^*M}| \approx 1 - \mu_{M\alpha} t^{-(1 - \mu_{M\alpha})(\ln r_M - \ln r_\alpha)/(\ln r_M - \ln r_m)}$$

where α is the index that maximizes the exponent (e.g., minimizes its absolute value). We note that the sum in which $r_i < r_j$ is empty when $i = M = 0$. We note further that the only constraint on α is that $\alpha > M = 0$. We now reconsider (4c), which is

$$\begin{aligned} &|p_{k^*k^*} - p_{Mk^*}| \\ &\quad \approx 1 - \sum_{j: r_k < r_j < r_m} \mu_{k^*j} - \mu_{k^*\beta} t^{-(1 - \mu_{k^*\beta})(\ln r_{k^*} - \ln r_\beta)/(\ln r_M - \ln r_m)} \\ &\qquad + \mu_{k^*\gamma} t^{-\mu_{k^*\gamma}(\ln r_\gamma - \ln r_{k^*})/(\ln r_M - \ln r_m)} \\ &\qquad - \mu_{Mk^*} t^{-(1 - \mu_{Mk^*})(\ln r_M - \ln r_{k^*})/(\ln r_M - \ln r_m)} \end{aligned}$$

where β is the index smaller than k^* that maximizes the exponent in which it appears and γ is the index greater than k^* that maximizes the exponent in which *it* appears. We note that the exponent of the last term must be less than or equal to the exponent in (46) and that both terms containing these exponents have the same sign. Now, we reconsider (4a). We see that

$$\begin{aligned} &\sum_{j: r_M, r_{k^*} > r_j} |p_{Mj} - p_{k^*j}| \\ &\quad \approx O(t^{-\min((1 - u_{M\delta})(\ln r_M - \ln r_\delta)/(\ln r_M - \ln r_m),\ (1 - \mu_{k^*\beta})(\ln r_{k^*} - \ln r_\beta)/(\ln r_M - \ln r_m))}) \end{aligned}$$

Once again, $\delta < k^*$ is chosen to minimize the appropriate exponent. The index β reappears because we need to maximize the same exponent as before. Thus, this sum eliminates at most one of the terms encountered previously. Upon reconsidering (4.4), we see that

$$\begin{aligned} &\sum_{j: r_i > r_j > r_{k^*}} |p_{Mj} - p_{k^*j}| \\ &\quad \approx \sum_{j: r_i > r_j > r_{k^*}} \mu_{k^*j} - \mu_{M\varepsilon} t^{(1 - \mu_{m\varepsilon})(\ln r_M - \ln r_\varepsilon)/(\ln r_M - \ln r_m)} \\ &\qquad - u_{k^*\gamma} t^{-u_{k^*}\gamma(\ln r_\gamma - \ln r_{k^*})/(\ln r_M - \ln r_m)} \end{aligned}$$

Again, we repeat the index γ to indicate that a term of this order has appeared before. Because this term appeared previously with the opposite sign, this second appearance cancels the first. It is clear, however, that that some terms with exponents between zero and unity remain in (3) under all circumstances. This is also true if $k^*=1$. Indeed, the only difference between this last case and the case considered previously is that the last sum considered is empty. Thus, (1) holds, as claimed.

To establish the second part of the theorem, we must show that C can be chosen such that, for $N(t) > C \ln t/(\ln r_M - \ln r_m)$,

$$\sum_{k=0}^{\infty} \alpha\left(\prod_{t_k=t}^{t_{k+1}} \mathbf{P}_{N(t)}\right) < \infty$$

for all choices of the sequence $0 \leqslant t_0 < t_1 < t_2 \cdots < t_k \cdots$. Recalling that

$$\alpha(\mathbf{P}_{N(t)}) = -\frac{1}{2}\max_{i,k}\left(-\mu_{ik} - \sum_{j:r_j<r_j} [(\mu_{ij}+\mu_{kj}) - |\mu_{ij}-\mu_{kj}|] + O(t^{-\lambda})\right)$$

we obtain an upper bound on the ergodic coefficient by choosing $i=M$, to obtain

$$\alpha(\mathbf{P}_{N(t)}) \leqslant \frac{1}{2}\min_k [\mu_{kM} + O(t^{-\lambda})]$$

A calculation similar to the one given above shows that, if $N(t) > C \ln t/(\ln r_M - \ln r_m)$ and $\min_k \mu_{kM} = 0$, then

$$\mathbf{P}_{N(t)} = \mathbf{P}_{N(\infty)} + O(t^{-C\lambda})$$

where λ is defined as before. We then have

$$\prod_{t_k=t}^{t_{k+1}} \mathbf{P}_{N(t)} = \mathbf{P}_{N(\infty)}^{t_{k+1}} + O(t^{-C\lambda})$$

Because $\mathbf{P}_{N(\infty)}$ is lower triangular, all of its powers are lower triangular. Hence,

$$\alpha\left(\prod_{t_k=t}^{t_{k+1}} \mathbf{P}_{N(t)}\right) = O(t_k^{-C\lambda})$$

We conclude that

$$\sum_{k=1}^{\infty} \alpha\left(\prod_{t_k=t}^{t_{k+1}} \mathbf{P}_{N(t)}\right) < \infty$$

for any choice of sequences $0 \leqslant t_0 < t_1 < t_2 \cdots < t_k < \cdots$ if $C > 1/\lambda$. ▌

So far, we have established a criterion under which the process will "forget" its initial state. We now ask whether the process converges to a steady-state distribution. More formally, we consider an arbitrary probability distribution **f** over the set $\mathscr{S}$ of possible states, and form vector/matrix products,

$$\mathbf{f} \prod_{t=0}^{t_1} \mathbf{P}_{N(t)} = \mathbf{f}^{(t_1)}$$

If there exists some probability vector $\mathbf{\Pi}$ such that

$$\lim_{t_1 \to \infty} \|\mathbf{f}^{(t_1)} - \mathbf{\Pi}\| = 0$$

for all starting vectors **f**, then the inhomogeneous Markov chain $\mathscr{M}$ defined by the matricies $\{\mathbf{P}_{N(t)}\}$ is said to be *strongly ergodic.* We note that weak ergodicity is a necessary, but not sufficient condition for strong ergodicity.

Our results for strong ergodicity are more dependent on the particular values of the μ_{ij} than our results for weak ergodicity. In particular, the strong ergodicity of $\mathscr{M}$ and the limiting probability distribution over states in $\mathscr{S}$ is dependent on whether $\boldsymbol{P}_{N(\infty)}$ is ergodic. Our first result in this direction is the following:

If $\mathbf{P}_{N(\infty)}$ is ergodic, then the Markov chain defined by the sequence of transition matrices $\{\mathbf{P}_{N(t)}\}$ is strongly ergodic, and, as $t \to \infty$, the state with the largest reproductive rate will be occupied with probability 1.

The first part of this result follows easily from a standard result in the theory of nonstationary Markov chains[10]:

If $\mathbf{P}_n$ is a sequence of finite-dimensional stochastic matrices such that $\|\mathbf{P}_n - \mathbf{P} \to 0\|$ as $n \to \infty$ and if $\mathbf{P}$ is ergodic, then the Markov chain defined by the sequence of transition matrices $\{\mathbf{P}_n\}$ is strongly ergodic.

In order to complete the proof of the theorem, we quote another standard result used in proving the theorem given above:

If $\{\mathbf{P}_n\}$ is a sequence of stochastic matrices such that $\|\mathbf{P}_n - \mathbf{P}\| \to 0$ as $n \to \infty$, then, for each positive integer k,

$$\|\mathbf{P}_{n+1}\mathbf{P}_{n+2}\mathbf{P}_{n+3} \cdots \mathbf{P}_{n+k} - \mathbf{P}\| \to 0$$

as $n \to 0$.

We combine these results to conclude that, for any $\varepsilon > 0$, we can find n and k large enough so that

$$\begin{aligned} \varepsilon &> \|\mathbf{P}_{n+1}\mathbf{P}_{n+2} \cdots \mathbf{P}_{n+k} - \mathbf{P}^k\| + \|\mathbf{P}^k - \mathbf{\Phi}\| \\ &\geqslant \|\mathbf{P}_{n+1}\mathbf{P}_{n+2} \cdots \mathbf{P}_{n+k} - \mathbf{\Phi}\| \end{aligned}$$

where $\mathbf{\Phi} = \lim_{k \to \infty} \mathbf{P}^k$. Because $\mathbf{P} = \mathbf{P}_{N(\infty)}$ is lower triangular and ergodic, a trivial calculation shows that the row vectors of $\mathbf{\Phi}$ are $(1, 0, 0, ..., 0)$, as claimed.

We turn next to the case where $\mathbf{P}_{N(\infty)}$ fails to be ergodic. We state our final (and most surprising) result:

> If $\mathbf{P}_{N(\infty)}$ is not ergodic, then the Markov chain defined by the sequence of transition matrices $\{\mathbf{P}_{N(t)}\}$ is strongly ergodic if $N(t) \leqslant \ln t/(\ln r_M - \ln r_m)$, but the stationary distribution does not necessarily assign positive probability to the state with the largest reproductive rate.

The proof of this result proceeds by choosing one of the relative maximum states s^* and partitioning the set of all possible states $\mathscr{S}$ into three subsets, the singleton set consisting only of s^* and the sets $\mathscr{S}^+$ and $\mathscr{S}^-$ defined by

$$\mathscr{S}^+ = \{s \in \mathscr{S} \mid r_s > r_{s^*}\}, \qquad \mathscr{S}^- = \{s \in \mathscr{S} \mid r_s < r_{s^*}\}$$

This choice of partition allows us to assume that the Markov chain under discussion has only three states: all elements of $\mathscr{S}^+$ ("state 1"), the state s^* ("state 2"), and all elements of $\mathscr{S}^-$ ("state 3"). We make a slight abuse of notation and redefine μ_{ij} as the rate of mutation from state i in the newly defined Markov chain to state j. We abuse notation again by redefining r_i to be the average reproduction rate for each type of organism in state i. We should compute the average at time t by using the probability distribution

$$\mathbf{f}\mathbf{P}_{N(t_0)})P_{N(t_0+1)} \cdots \mathbf{P}_{N(t)}\, d\mathscr{P}\{\mathbf{f}\}$$

where $d\mathscr{P}\{\mathbf{f}\}$ is some *a priori* distribution of starting states $\mathbf{f}$. Fortunately, though, this complication is irrelevant to the thrust of the argument, because the average r for $\mathscr{S}^+$ and $\mathscr{S}^-$ is always between the largest and smallest r's associated with these states. A similar comment applies to the μ's.

For notational convenience, we introduce the symbols

$$\lambda_{12} = (1 - \mu_{12}) \frac{\ln r_1 - \ln r_2}{\ln r_1 - \ln r_3}, \quad \lambda_{13} = (1 - \mu_{13})$$

$$\lambda_{23} = (1 - \mu_{23}) \frac{\ln r_2 - \ln r_3}{\ln r_1 - \ln r_3}, \quad \lambda_{31} = \mu_{31}, \quad \lambda_{32} = \mu_{32} \frac{\ln r_2 - \ln r_3}{\ln r_1 - \ln r_3}$$

We can then repeat the computations given previously and we conclude that, to first order in all terms,

$$\mathbf{P}_{N(t)=} \begin{pmatrix} 1 - \mu_{12} t^{-\lambda_{12}} - \mu_{13} t^{-\lambda_{13}} & \mu_{12} t^{-\lambda_{12}} & \mu_{13} t^{-\lambda_{13}} \\ 0 & 1 - \mu_{23} t^{-\lambda_{23}} & \mu_{23} t^{-\lambda_{23}} \\ \mu_{31} - \mu_{31} t^{-\lambda_{31}} & \mu_{32} - \mu_{32} t^{-\lambda_{32}} & 1 - \mu_{31} - \mu_{32} + \mu_{31} t^{-\lambda_{31}} + \mu_{32} t^{-\lambda_{32}} \end{pmatrix}$$

We show that this sequence of transition matrices is strongly ergodic by an application of the following theorem[11]:

Let $\mathbf{P}_{N(t)}$ be the one-step transition matrix of a discrete-time, nonstationary, finite Markov chain composed of C states. Suppose that

$$\lim_{t \to \infty} \mathbf{P}_{N(t)} = \mathbf{P}_\infty \qquad \text{and} \qquad \mathbf{P}_{N(t)} = \mathbf{P}_\infty + \mathbf{V}(t)$$

where $\mathbf{P}_\infty$ is a constant matrix. Suppose further that $\mathbf{P}_\infty$ has exactly two ergodic, aperiodic components and a positive, but otherwise arbitrary number of transient states. Let $\mathbf{m}^{(1)}$ and $\mathbf{m}^{(2)}$ be the two eigenvectors that satisfy

$$\mathbf{m}\mathbf{P}_\infty = \mathbf{m}$$

which, by hypothesis, must be of the form

$$\mathbf{m}^{(1)} = (m_1^{(1)}, m_2^{(1)}, \ldots, m_{u_1}^{(1)}, \underbrace{0, \ldots, 0}_{C - \mu_1 \text{ repetitions}})$$

and

$$\mathbf{m}^{(2)} = (\underbrace{0, \ldots, 0}_{u_1 \text{ repetitions}}, m_{u_1+1}^{(2)}, m_{u_1+2}^{(2)}, \ldots, m_{u_1+u_2}^{(2)}, \underbrace{0, \ldots, 0}_{C - u_1 - u_2 \text{ repetitions}})$$

Let $\mathbf{\Pi}(t)$ satisfy $\mathbf{\Pi}(t)\,\mathbf{P}_{N(t)} = \mathbf{\Pi}(t)$, let

$$\begin{aligned}\phi(t) = \sum_{i=1}^{u_2} m_i^{(2)}[&V_{u_1+i,1}(t) + \cdots + V_{u_1+i,u_1(t)} \\ &+ V_{u_1+i,u_1+u_2+1}(t)\, z_{u_1+u_2+1,\,1} + \cdots + V_{u_1+i,C}(t)\, z_{n,1}]\end{aligned}$$

and let

$$\begin{aligned}\psi(t) = \sum_{i=1}^{u_2} m_i^{(1)}[&V_{i,u_1+1}(t) + \cdots + V_{i,u_1+u_2(t)} \\ &+ V_{i,u_1+u_2+1}(t)\, z_{u_1+u_2+1,2} + \cdots + V_{i,C}(t)\, z_{C,2}]\end{aligned}$$

where $z_{jk} = \Pr\{\text{eventual transition from transient state } j \text{ to ergodic component } k\}$. If $\sum_{i+1}^{\infty} [\phi(t) + \psi(t)] = \infty$ and $\lim_{t\to\infty} \mathbf{\Pi}(t)$ exists, then $\lim_{t\to\infty} \prod_{k=1}^{t} \mathbf{P}_{N(k)}$ exists and is equal to the matrix whose rows are all $\lim_{t\to\infty} \mathbf{\Pi}(t)$.

We have

$$\mathbf{P}_\infty = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ u_{31} & \mu_{32} & 1-\mu_{31}-\mu_{32} \end{pmatrix}$$

and

$$\mathbf{V}(t) = \begin{pmatrix} -u_{12}t^{-\lambda_{12}} - \mu_{13}t^{-\lambda_{13}} & \mu_{12}t^{-\lambda_{12}} & \mu_{13}t^{-\lambda_{13}} \\ 0 & -\mu_{23}t^{-\lambda_{23}} & \mu_{23}t^{-\lambda_{23}} \\ -\mu_{31}t^{-\lambda_{31}} & -\mu_{32}t^{-\lambda_{32}} & \mu_{31}t^{-\lambda_{31}} + \mu_{32}t^{-\lambda_{32}} \end{pmatrix}$$

It is then clear that $\mathbf{m}^{(1)} = (1, 0, 0)$ and that $\mathbf{m}^{(2)} = (0, 1, 0)$. It is also easy to check that the ith component of $\mathbf{\Pi}$ is $\phi_i/(\phi_1 + \phi_2 + \phi_3)$, where

$$\phi_1 = \frac{\mu_{31}(1 - t^{-\lambda_{31}})}{\mu_{12}t^{-\lambda_{12}} + \mu_{13}t^{-\lambda_{13}}}$$

$$\phi_2 = \mu_{12}\frac{\mu_{31}}{\mu_{23}}\frac{(1 - t^{-\lambda_{31}})\, t^{-\lambda_{12}+\lambda_{23}}}{\mu_{12}t^{-\lambda_{12}} + \mu_{13}t^{-\lambda_{13}}} + \frac{\mu_{32}}{\mu_{23}}(1 - t^{-\lambda_{32}})\, t^{\lambda_{23}}$$

$$\phi_3 = 1$$

We conclude that $\lim_{t\to\infty} \mathbf{\Pi}$ always exists, but that it is either $(1, 0, 0)$ or $(0, 1, 0)$, depending on whether $\min(\lambda_{12}, \lambda_{13})$ is greater than or less than λ_{23}.

Next, we verify that the rest of the conditions of the theorem hold, so that we can conclude that $\lim_{t\to\infty} \mathbf{\Pi}(t)$ is, in fact, the limiting distribution. After an easy computation, we see that

$$z_{31} = \frac{\mu_{31}}{\mu_{31} + \mu_{32}}, \qquad z_{32} = \frac{\mu_{32}}{\mu_{31} + \mu_{32}}$$

$$\phi(t) = V_{2,1} + V_{2,3}z_{3,1} = \frac{\mu_{23}\mu_{31}}{\mu_{32} + \mu_{32}}\, t^{-\lambda_{23}}$$

and

$$\psi(t) = V_{1,2} + V_{1,3}z_{3,2} = \mu_{12}t^{-\mu_{12}} + \frac{\mu_{13}\mu_{32}}{\mu_{31} + \mu_{32}}\, t^{-\lambda_{13}}$$

It follows that

$$\sum_{t=0}^{\infty} [\phi(t) + \psi(t)] = \infty$$

unless $\mu_{12} = \min(\mu_{13}, \mu_{32}) = 0$. ∎

However, this model does have "more or less" optimizing behavior for physically meaningful values of the parameters μ and r. Recalling that r_1 is an average of all states in $\mathscr{S}^+$ and that r_3 is an average over all states in $\mathscr{S}^-$ and assuming that $\mathscr{S}^+$ and $\mathscr{S}^-$ have no local maxima, then we know that $r_1 \to r_M$ and $r_3 \to r_{s_-} = \max_{s \in \mathscr{S}^-} r_s$. Presumably, the r values for the different states are relatively closely and evenly spaced, so that $\ln r_M - \ln r_2 > \ln r_2 - \ln r_{s_-}$. It will then be the case that $\lambda_{12} > \lambda_{23}$, and the stationary distribution is $(1, 0, 0)$. If either $\mathscr{S}^+$ or $\mathscr{S}^-$ has local maxima, we can apply the same line of argument to each local maximum in succession. It is in this sense that optimizing behavior occurs.

5. CONCLUSIONS

The obvious question to ask at this point is what these calculations have to do with real biological systems. While the present model is clearly an oversimplification, "historical accidents," reminiscent of the frozen-in crystal defects that arise from quenching, certainly abound in biological structures. Among the examples that come to mind are the human appendix, atavisms such as the "hen's teeth and horses toes" of Gould's book,[12] and the universality of the genetic code and basic biochemistry. Other examples are given in Ref. 1.

Although our results are strictly true only in the limit as t becomes infinite, they should also be good approximations when t is large, provided that the basic assumptions of the model are not violated. Another context in which these assumptions might be justified is if the population is one of many weakly interacting populations residing at single points on a spatially extended domain. The foregoing analysis then describes the entire ensemble of populations in the domain. If we assume further that each population is of roughly the same size $N(t)$, then it is indeed plausible that number of mutants or migrants is proportional to $N(t)$, as posited before. Also note that, if $r_s - r_{s'} = O(1)$ then the birth-and-death process describing the struggle for survival of the mutant is essentially deterministic, expontial growth. We expect, therefore, that the time required for either the wild type or the mutant to die out is usually $O(\ln N)$. This observation lends support to the use of two time scales in the analysis.

The results are reminiscent of the results obtained by the introduction of diffusion in chemical kinetics. When the diffusion proceeds rapidly relative to the other processes in the system (i.e., the diffusion coefficient is relatively large), then the uniform distribution is the unique stationary distribution. If the diffusion coefficient is made sufficiently small, there may be multiple steady-state solutions to the reaction-diffusion equations governing the system. Because high temperatures facilitate diffusion, this observation is further evidence that $N(t)$ in evolving biological systems and inverse temperature in physical systems play a similar role.

ACKNOWLEDGMENTS

I thank J. K. Percus, Alan Sokal, and Steve R. White for useful discussions.

REFERENCES

1. S. J. Gould, *The Panda's Thumb: More Reflections in Natural History* (Norton, New York, 1980).
2. E. Weinberger, Ph. D. Thesis, Courant Institute of Mathematical Sciences, New York (1987).
3. M. Eigen and P. Schuster, *The Hypercycle: A Principle of Natural Self-Organization* (Springer, New York, 1979).
4. J. Gillespie, *Evolution* **38**(5):1116–1129 (1984).
5. S. Karlin and H. Taylor, *A First Course in Stochastic Processes* (Academic Press, New York, 1981).
6. W. Feller, *Introduction to Probability Theory and Its Applications* (Wiley, New York, 1968).
7. S. Karlin and H. Taylor, *A Second Course in Stochastic Processes* (Academic Press, New York, 1981).
8. S. Kirkpatrick, C. D. Gelatt, Jr. and M. P. Vecci, *Annealing*, *Science* **220**:671–680 (1983).
9. E. H. L. Aarts and P. J. M. van Laarhoven, *Philips J. Res.* **40**:193–226 (1985).
10. D. Isaacson and R. Madsen, *Markov Chains: Theory and Applications* (Wiley, and Sons, New York, 1976).
11. Basilis Gidas, *J. Stat. Phys.* **39**:73–131 (1985).
12. S. J. Gould, *Hen's Teeth and Horses Toes* (Norton, New York, 1983).