Journal of Statistical Physics, Vol. 97, Nos. 3/4, 1999

Analysis of a Population Genetics Model with Mutation, Selection, and Pleiotropy

S. N. Coppersmith,[1] Robert D. Blank,[2] and Leo P. Kadanoff[1]

Received March 28, 1999; final March 28, 1999

We investigate a population genetics model introduced by Waxman and Peck[(1)] incorporating mutation, selection, and pleiotropy (one gene affecting several unrelated traits). The population is infinite, and continuous variation of genotype is allowed. Nonetheless, Waxman and Peck showed that if each gene affects three or more traits, then the steady-state solution of the model can have a nonzero fraction of the population with identical alleles. We use a recursion technique to calculate the distribution of alleles at finite times as well as in the infinite-time limit. We map Waxman and Peck's model into the mean-field theory for Bose condensation, and a variant of the model onto a bound-state problem in quantum theory. These mappings aid in delineating the region of parameter space in which the unique genotype occurs. We also discuss our attempts to correlate the statistics of DNA-sequence variation with the degree of pleiotropy of various genes.

KEY WORDS: Population genetics; pleiotropy; quantum mechanics.

1. INTRODUCTION

Recently, Waxman and Peck[(1)] introduced a simple population genetics model of an infinite population with a continuous distributions of alleles[3] incorporating pleiotropy (one gene affecting several characters of an organism). They demonstrated that when the number of characters affected is greater than two, the long-time steady state solution of their model can

[1] The James Franck Institute, University of Chicago, Chicago, Illinois 60637.

[2] The Hospital for Special Surgery, New York, New York 10021, and Department of Medicine, Cornell University Medical College, New York, New York.

[3] Waxman and Peck's model is based on the continuum-of-alleles model of Kimura.[(2)] Other important investigations of pleiotropic models include lande,[(3)] Turelli,[(4)] Wagner,[(5)] and Gavrilets and de Jong.[(6)]

0022-4715/99/1100-0429$16.00/0

have a nonzero fraction of the population with identical alleles, and that this phenomenon does not occur if the number of characters affected per gene is two or fewer.

In this paper we study the time evolution of the distribution of alleles in this model. Initially, the distribution of alleles is continuous, and at infinite time there is an infinitely narrow (δ-function) peak; we calculate the distribution of alleles at long but finite times by using an expansion in a sum of functions (Gaussians) particularly picked to meet the requirements of this problem. We also examine the relationship between the model with discrete generations used by Waxman and Peck and a continuous-time model in which both mutation and selection occur continuously. We demonstrate that the qualitative behavior of the two models is the same, but that the dependence of the behavior on the parameters can be different in the two models.

We show that the population genetics models we consider can be mapped onto problems in quantum mechanics, specifically Bose–Einstein condensation[7] and motion of a particle in a central potential.[8, 9] The mapping onto Bose condensation is performed on the discrete-time model in the limit that selection is very strong, so that only organisms with fitnesses very near optimum survive each generation. The mapping onto the quantum particle in a central potential applies more generally. The simplest case has the potential independent of time. This case corresponds to the a limit of the population model in which selection and mutation both occur continuously. The close relation between the continuous-time population biology models and Schrödinger's equation has been exploited by Bürger[4] to obtain many general results on the long-time behavior of models of the type studied here. In this paper, we focus on the mapping for a specific model, which yields a simple physical interpretation of the emergence of a unique genotype, and allows the extraction of the time-dependence of the fitness peak as well as comparison with the discrete-time model. We find that the behavior is qualitatively identical when mutation is continuous and when it is discrete, though there are some quantitative differences.

We also discuss our attempts to relate the results from this model to the statistical properties of the sequences that are included in various genetic databases. We attempt to correlate observed sequence variations with estimates of the degree of pleiotropy of a sample of genes. The results of these attempts are inconclusive.

This paper is organized as follows. In Section 2 we introduce the theoretical models. Section 3 presents our analysis of the time evolution of the discrete version of the model, demonstrating that the behavior can be

[4] A recent paper with references to previous work is ref. 10.

extracted analytically in a limit in which the typical jump caused by mutation is very large. In Section 4 we present our analysis of the continuous-time model. Section 5 shows that the discrete-time model can be mapped onto a continuous-time quantum mechanics problem of a particle in the presence of a time-dependent potential. This mapping is used to demonstrate that the behavior in the limit of infinite time is the same for all initial conditions. Section 6 presents our examination of DNA sequence archives and our attempts to correlate the statistics of documented sequence variations to the degree of pleiotropy of various genes. Section 7 recapitulates our main conclusions.

2. THE MODELS

One model we study is that of Waxman and Peck.[1] In this model, the population is infinite, the phenotypic variation is continuous, each gene affects N uncorrelated characters, and the effects of different genes are uncorrelated (no linkage disequilibrium or epistasis). The model assumes a very large population of haploid and asexual organisms with discrete generations. Let $\vec{x}$ be a continuous vector with N components which represents characters that determine the viability of an organism, and $\phi(\vec{x}, t)$ be the normalized probability density that an organism in the population has the characters given by $\vec{x}$ at time t. The normalization condition for the probability is given by an N-dimensional integral as

$$\int \phi(\vec{x}, t)\, d^N x = 1 \tag{1}$$

At each generation an organism with characters $\vec{x}$ survives viability selection with a probability proportional to $\exp(-\vec{x}^2/2V_s)$ (this is the definition of V_s). This Gaussian selection will play an essential role in our analysis of the model. After this selection, a fraction θ of the population mutates; it is assumed that if a mutation occurs, then the probability that the mutant takes on the value $\vec{x}$ given the parental value $\vec{y}$ is $f(\vec{x}-\vec{y})$.

Since $\phi(\vec{x}, t)$ is a probability density, it is normalized at every time t. Thus, for this model, the equation for the evolution of $\phi(\vec{x}, t)$ is

$$\bar{w}(t+1)\, \phi(\vec{x}, t+1) = (1-\theta)\, w_1(\vec{x})\, \phi(\vec{x}, t) + \theta \int f(\vec{x}-\vec{y})\, w_1(\vec{y})\, \phi(\vec{y}, t)\, d^N y \tag{2}$$

where the fitness factor $w_1(\vec{x})$ is

$$w_1(\vec{x}) = \exp[-\vec{x}^2/2V_s] \tag{3}$$

The multiplicative factor $\bar{w}(t+1)$ in Eq. (2) ensures that the probability density $\phi(\vec{x}, t+1)$ is normalized to unity at every time step. Integrating Eq. (2) over all $\vec{x}$ gives the Waxman and Peck result for this normalization:

$$\bar{w}(t+1) = \int \phi(\vec{x}, t)\, w_1(\vec{x})\, d^N x \tag{4}$$

We follow Waxman and Peck and use a Gaussian function for the mutation probability:

$$f(\vec{x} - \vec{y}) = (2\pi m^2)^{-N/2} \exp[-(\vec{x} - \vec{y})^2/2m^2] \tag{5}$$

where m^2 describes the variance of the mutant effects for a single character.

We will also consider a model in which both mutation and selection occur continuously in time. The continuous time model can be obtained from the discrete-time one via a limiting process in which τ, the time between successive generations, approaches zero.[11] To see this explicitly, we define the continuous-time variable $T = \tau t$ (again, τ is the interval between generations; t is the generation number), and we posit that the parameters θ and V_s describing mutation and selection scale with τ as $\theta = \tau\Theta$, and $V_s = V/\tau$, with Θ and V independent of τ. With this scaling, the change in the population distribution during each generation vanishes while the change per unit time can tend to a nonzero limit.

Defining $\Phi(\vec{x}, T) = \phi(\vec{x}, t)$ and $\bar{w}(T) = \bar{w}(t)$, we find, as $\tau \to 0$,

$$w_1(\vec{x}) = \exp\left[-\frac{\tau \vec{x}^2}{2V} \right] = 1 - \frac{\tau \vec{x}^2}{2V} + \mathcal{O}(\tau^2) \tag{6}$$

and

$$\bar{w}(T+\tau) = 1 - \frac{\tau \vec{x}^2}{2V} V_{\Phi}(T) + \mathcal{O}(\tau^2) \tag{7}$$

where

$$V_{\Phi}(T) = \int \vec{x}^2 \Phi(\vec{x}, T)\, d^N x \tag{8}$$

Expanding Eq. (2) for small τ, rearranging terms, and taking the limit $\tau \to 0$, one finds

$$\frac{\partial \Phi(\vec{x}, T)}{\partial T} = \left[a(T) - \frac{\vec{x}^2}{2V} \right] \Phi(\vec{x}, T) + \Theta \int d^N y f(\vec{x} - \vec{y})[\Phi(\vec{y}, T) - \Phi(\vec{x}, T)] \tag{9}$$

where $a(T) = V_{\Phi}(T)/2V$.

Note that in these models in order to have high fitness, each of the N components of $\vec{x}$ must be near zero. Since each mutation of a gene alters N characters, pleiotropy reduces the probability that a newly mutated individual is highly fit.

In the next two sections we calculate the evolution of the probability density for different N, with an emphasis on the behavior near $\vec{x}=0$ at long times. In contrast to Waxman and Peck,[1] who solve approximately the fixed point equation to obtain a time-independent probability density, we demonstrate that there is only one fixed point solution, and calculate the time-dependence of the approach to it. In Section 3 we examine the discrete-time model Eq. (2); Section 4 discusses the continuous-time equation Eq. (9).

3. TIME EVOLUTION OF DISCRETE EQUATIONS

In this section we consider the discrete-time evolution Eq. (2). We will determine its behavior for Gaussian initial conditions in the strong selection limit $V_s/m^2 \ll 1$, in which the analysis simplifies considerably. Though typical biological systems are not described by the strong selection limit,[12] below in Section 5 we will argue that the behavior is not sensitive either to the use of this limit or to the choice of initial condition.

Our method of solution of Eq. (2) exploits the following observation. Suppose at some time t, $\phi(\vec{x}, t)$ is a normalized Gaussian with variance α, so that $\phi(\vec{x}, t) = G_\alpha(\vec{x})$, where[5]

$$G_\alpha(\vec{x}) = (2\pi\alpha)^{-N/2} \exp\left[-\frac{\vec{x}^2}{2\alpha} \right] \tag{10}$$

Then $\phi(\vec{x}, t+1)$ is the sum of two Gaussians. This result follows because Eq. (2) involves the processes of multiplication and of convolution with a Gaussian. The product of two Gaussians is

$$G_\alpha(\vec{x})\, G_\beta(\vec{x}) = \left(\frac{\gamma}{2\pi\alpha\beta} \right)^{N/2} G_\gamma(\vec{x}), \qquad \text{where} \quad 1/\gamma = 1/\alpha + 1/\beta \tag{11}$$

while the convolution of two Gaussians is

$$\int G_\alpha(\vec{x} - \vec{y})\, G_\beta(\vec{y})\, d^N y = G_\delta(\vec{x}), \qquad \text{where} \quad \delta = \alpha + \beta \tag{12}$$

[5] Strictly speaking, α is the variance of each component of $\mathbf{x}$.

Therefore, if $\phi(\vec{x}, 0)$ is a Gaussian, $\phi(\vec{x}, 0) = G_{\alpha}(\vec{x})$, then $\phi(\vec{x}, 1)$ is the sum of two Gaussians,

$$\phi(\vec{x}, 1) = (1-\theta)\, G_{\beta_1(\alpha)}(\vec{x}) + \theta G_{\beta_2(\alpha)}(\vec{x}) \tag{13}$$

with the two β's given by

$$\beta_1(\alpha) = \frac{\alpha}{1+\alpha/V_s} \tag{14}$$

$$\beta_2(\alpha) = m^2 + \beta_1(\alpha) \tag{15}$$

It follows immediately that if $\phi(\vec{x}, t)$ is the sum of finitely many Gaussians, then $\phi(\vec{x}, t+1)$ is also the sum of finitely many Gaussians. Instead of doing integrals at each time step, we can write recursion relations for the evolution of the widths and amplitudes of the Gaussians. Our method of solution has similarities with that used by Kingman[13] on a simpler model.

Explicitly, we write $\phi(\vec{x}, t)$ in the form

$$\phi(\vec{x}, t) = \sum_i A_i(t)\, G_{\alpha_i(t)}(\vec{x}) \tag{16}$$

with the normalization condition $\sum_i A_i = 1$. Then $\phi(\vec{x}, t+1)$ must satisfy

$$\bar{w}(t+1)\, \phi(\vec{x}, t+1) = (1-\theta) \sum_i B_i(t)\, G_{\beta_1(\alpha_i(t))}(\vec{x}) + \theta \sum_i B_i(t)\, G_{\beta_2(\alpha_i(t))}(\vec{x}) \tag{17}$$

with

$$B_i(t) = A_i(t)[\,1 + \alpha_i(t)/V_s\,]^{-N/2} \tag{18}$$

$$\bar{w}(t+1) = \sum_i B_i(t) \tag{19}$$

Equations (17)–(19) simplify considerably when $V_s/m^2 \ll 1$. In this case, the standard deviation of the mutation probability function, f, is much greater than the width of the population distribution after selection. In this big-mutation-jump limit, $\beta_1(m^2) \approx V_s$, $\beta_2(m^2) \approx m^2$, $\beta_1(V_s/n) = V_s/(n+1)$, $\beta_2(V_s/n) \approx m^2$.[6] (Here, n is any positive integer.)

[6] Because the newly mutated population is a Gaussian of width determined by m^2 for any distribution, this approximation is equivalent to the "house-of-cards" approximation introduced by Kingman[13] and investigated in Turelli[12] and Barton and Turelli.[14]

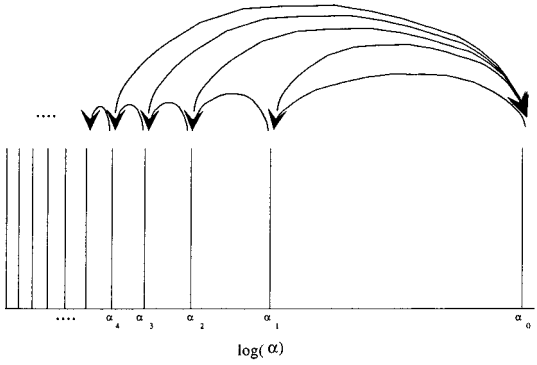

Fig. 1. Sketch of transitions between α_i's in the limit $V_s/m^2 \ll 1$. Each α_i corresponds to a Gaussian in $\phi(\vec{x}, t)$.

To analyze these recursion relations, first note that in the absence of mutations ($\theta = 0$), if $\phi(\vec{x}, 0) = G_{\alpha_0}(\vec{x})$, then $\phi(\vec{x}, t) = G_{\alpha(t)}(\vec{x})$, with

$$\alpha(t) = \frac{\alpha_0}{1 + \alpha_0 t/V_s} \tag{20}$$

Thus, the population distribution remains Gaussian, and at long times, its width scales as $|\vec{x}| \sim t^{-1/2}$. As expected, selection continually narrows the distribution.

Now we examine Eq. (2) with $\theta > 0$ in the strong selection limit $V_s/m^2 \ll 1$. Consider the time evolution of an initial distribution consisting of a Gaussian with variance $\alpha = m^2$. After one time step, the distribution splits into two Gaussians, an unmutated population with variance $\beta \approx V_s$ and a mutated population with variance $\beta \approx m^2$. After two time steps, the population consists of three Gaussians with β values m^2, V_s, and $V_s/2$. After n steps, the population consists of $n+1$ Gaussians with β values m^2, V_s, $V_s/2$,..., V_s/n. We define $\alpha_0 = m^2$ and $\alpha_i = V_s/i$ for $i \geqslant 1$. (Note that these α's are time-independent.) If we start with a Gaussian initial distribution of width much greater than m, then after n steps, the distribution consists of $n+1$ Gaussians with β values α_0,..., α_n. We describe the time evolution using rate equations for the populations in these states. Figure 1 is a sketch of the transitions that can occur between the different α's. Note that at a given time, a state α_i makes transitions to two states, α_0 and α_{i+1}.

Writing $\phi(\vec{x}, t)$ as

$$\phi(\vec{x}, t) = \sum_{i=0}^{t} A_i(t)\, G_{\alpha_i}(\vec{x}) \tag{21}$$

and specializing Eqs. (17)–(19) to the limit $V_s/m^2 \ll 1$, we find that the $A_i(t)$ must obey

$$A_0(t) = \theta, \qquad t > 0 \tag{22}$$

$$A_1(t) = \frac{(1-\theta)}{\bar{w}(t)} Q A_0(t-1), \qquad t > 0 \tag{23}$$

$$A_i(t) = \frac{1-\theta}{\bar{w}(t)} \left[\frac{i-1}{i} \right]^{N/2} A_{i-1}(t-1), \qquad i > 1, \quad t > 0 \tag{24}$$

and

$$\bar{w}(t+1) = Q A_0(t) + \sum_{i=1}^{t} A_i(t) \left(\frac{i}{i+1} \right)^{N/2} \tag{25}$$

We have defined $Q = (1 + m^2/V_s)^{-N/2}$; note that $Q \ll 1$ in the limit we consider.

First consider the first two steps of the evolution. Starting with the initial condition $A_0(0) = 1$, $A_i(0) = 0$ for $i \geqslant 1$, one finds

$$\begin{aligned} A_0(1) &= \theta \\ A_1(1) &= 1 - \theta \end{aligned} \tag{26}$$

So far not much has happened. But a key thing happens at the next step:

$$\begin{aligned} A_0(2) &= \theta \\ A_1(2) &= \frac{\theta Q (1-\theta)}{\theta Q + (1-\theta)\, 2^{-N/2}} \\ A_2(2) &= \frac{(1-\theta)^2\, 2^{-N/2}}{\theta Q + (1-\theta)\, 2^{-N/2}} \end{aligned} \tag{27}$$

Note that $A_1(2) \propto Q \ll 1$. In fact, for any $t > 2$, all A_i's with $0 < i < t$ are proportional to Q. However, as $t \to \infty$, the number of these terms diverges. So it is not obvious whether as $t \to \infty$ a solution exists where $A_0(t) = \theta$, $A_t(t) = \mathcal{O}(1)$, and all other $A_i(t)$'s are small, which would imply that the Gaussian describing the unmutated population contains a nonzero fraction of the total population. We will find that such a solution can exist only when $N > 2$.

3.1. Steady-State Solutions

First we find the steady-state solutions in the longtime limit $t \to \infty$. In steady state, the time arguments on the A_i and on $\bar{w}$ can be dropped, and the recursion relations become

$$A_0 = \theta \tag{28}$$

$$A_1 = \theta Q v, \tag{29}$$

$$A_i = v\left[\frac{i-1}{i}\right]^{N/2} A_{i-1} \qquad (i > 1) \tag{30}$$

and

$$\frac{1-\theta}{v} = \theta Q + \lim_{t\to\infty} \sum_{i=1}^{t} A_i \left(\frac{i}{i+1}\right)^{N/2} \tag{31}$$

where we have defined $v = (1-\theta)/\bar{w}$. We explicitly allow for the possibility that the amplitude $A_\infty \equiv \lim_{t\to\infty} A_t(t)$ is nonzero.

The solution to Eqs. (28)–(31) is

$$\begin{aligned} A_0 &= \theta \\ A_i &= \theta Q i^{-N/2} v^i, \qquad i \geqslant 1 \end{aligned} \tag{32}$$

where v must satisfy

$$\frac{1-\theta}{v} = A_\infty + \theta Q \sum_{i=0}^{\infty} v^i \left(\frac{1}{i+1}\right)^{N/2} \tag{33}$$

Note that $v \leqslant 1$, for otherwise the A_i cannot sum to unity, which is required for normalization of the probability.[7] Since as v increases the left-hand side of Eq. (33) monotonically decreases and the right-hand side monotonically increases, a solution with $A_\infty = 0$ can exist only if $\theta Q \sum_{i=0}^{\infty} [1/(i+1)]^{N/2} > 1-\theta$. Conversely, if this inequality is not satisfied, then we expect $A_\infty > 0$. Since $A_\infty \sim \lim_{t\to\infty} v^t$, we require $v = 1$ if A_∞ is nonzero. We show below that such a solution is the long-time limit of the solution of the time-dependent equations.

If $A_\infty \neq 0$, then a nonzero fraction of the population never mutates. The probability density of this subpopulation continually narrows, and as $t \to \infty$ the alleles in this subpopulation are identical. Conversely, if $A_\infty = 0$, then at long times the probability of observing an unmutated individual tends to zero, and the probability density retains a finite width at infinite time.

[7] This inequality is derived in Waxman and Peck,[1] and is closely related to results derived by Bürger and collaborators.[10, 15–17]

First consider the case when $A_\infty \neq 0$. Since the sum in Eq. (33) converges if $N > 2$, and diverges otherwise, we see that $A_\infty \neq 0$ can occur only if $N > 2$. When the sum does converge, the unmutated fitness peak, whose amplitude is A_∞ and whose width is given by Eq. (20), can contain a nonzero fraction of the total weight. Evaluating Eq. (31) with $v = 1$, we find $A_\infty = 1 - \theta - \theta Q \zeta(N/2)$, where $\zeta(N/2)$ is the Riemann zeta function.[18, 19]

Now we consider the possibility of a solution in which the unmutated population constitutes an infinitesimal fraction of the total population as $t \to \infty$. Since $A_\infty = 0$, we must have

$$\frac{1-\theta}{\theta Q} = L_{i_{N/2}}(v) \tag{34}$$

where $v < 1$ and $Li_a(x)$ is the polylogarithm function:[20]

$$Li_a(x) \equiv \sum_{i=1}^{\infty} \frac{1}{i^a} x^i \tag{35}$$

When $a > 1$, $Li_a(1) = \zeta(a)$ is finite, where again $\zeta(z)$ is the Riemann zeta function. Therefore, for $N > 2$, the right-hand side of Eq. (34) is bounded as $v \to 1$, and a solution exists only if $(1-\theta)/(\theta Q) < \zeta(N/2)$. In the $Q \ll 1$ limit that we have assumed, this happens only when $1 - \theta$ is very small also. If this condition is not satisfied, then a nonzero fraction of the population must be in the unmutated state.

Now we calculate the function $\phi(\vec{x}, t)$ as $t \to \infty$.[8] First we consider the case when ϕ has a δ-function piece.

In this regime, as $t \to \infty$, $A_0 = \theta$, $A_i = \theta Q (1/i)^{N/2}$ $(i \geqslant 1)$, and A_∞, the weight in the δ-function, is $A_\infty = 1 - \theta - \theta Q \zeta(N/2)$, where $\zeta(N/2)$ is the Riemann zeta function. Therefore, as $t \to \infty$,

$$\begin{aligned}
\phi(\vec{x}, t) &\to (1 - \theta - \theta Q \zeta(N/2))\, \delta(x) + \theta (2\pi m^2)^{-N/2} \exp\left[-\frac{\vec{x}^2}{2m^2} \right] \\
&\quad + \theta Q \sum_{n=1}^{\infty} (1/n)^{N/2} (2\pi V_s/n)^{-N/2} \exp\left[-\frac{n\vec{x}^2}{2V_s} \right] \\
&= (1 - \theta - \theta Q \zeta(N/2))\, \delta(x) + \theta (2\pi m^2)^{-N/2} \exp\left[-\frac{\vec{x}^2}{2m^2} \right] \\
&\quad + \theta Q (2\pi V_s)^{-N/2} \left(\exp\left[\frac{\vec{x}^2}{2V_s} \right] - 1 \right)^{-1}
\end{aligned} \tag{36}$$

[8] The form of the distribution calculated by Waxman and Peck (their footnote 29) does not apply here because they require simultaneous validity of the inequalities $\theta V_s/m^2 \ll 1$, $m^2/V_s \ll 1$, and $N \ll V_s/m^2$. Our calculation applies whenever $V_s/m^2 \ll 1$.

Thus, we see that when $\phi(\vec{x}, t)$ has a δ-function piece as $t \to \infty$, there is in addition a divergent contribution at small x, proportional to $1/\vec{x}^2$. When the distribution is smooth, we have $A_0 = \theta$ and $A_i = \theta Q(1/i)^{N/2} v^i$ $(i \geqslant 1)$, and $v < 1$, and we obtain as $t \to \infty$

$$\phi(\vec{x}, t) \to \theta(2\pi m^2)^{-N/2} \exp\left[-\frac{\vec{x}^2}{2m^2} \right] + \theta Q(2\pi V_s)^{-N/2} \left(v^{-1} \exp\left[\frac{\vec{x}^2}{2V_s} \right] - 1 \right)^{-1} \tag{37}$$

The sums that arise here are identical to those that come up in the calculation of Bose–Einstein condensation for an ideal Bose gas.[7] The weight in the δ-function in the genotype distribution, A_∞, is analogous to the condensate fraction in the Bose condensation problem. The parameter N, which here describes the number of traits affected by a mutation, is the number of spatial dimensions in the Bose gas calculation. That the superfluid fraction must be zero for a Bose gas in one and two dimensions is a special case of a general result in the theory of phase transitions.[21–23]

3.2. Time-Dependent Solutions

We now discuss the time evolution of the population distribution. We begin with some qualitative remarks. When $v < 1$, the sum in the solution Eq. (33) converges geometrically. Therefore, in this regime one expects the approach to the $t \to \infty$ limit to be exponential. In the regime where a δ-function contribution is present as $t \to \infty$, the δ-function is the long-time limit of a Gaussian describing the unmutated population. The variance of this Gaussian narrows as $1/t$, so we expect the approach to the long-time limit to be a power law. These expectations are supported by the explicit calculation that we now present. We also show that the long-time corrections to the amplitude of the δ-function are also proportional to $1/t$.

We again write recursion relations describing the transitions between the various Gaussians, now allowing for explicit time dependence in the $A_i(t)$. Defining $v(t) = (1-\theta)/\bar{w}(t+1)$, these recursion relations are, for $t > 0$,

$$A_0(t) = \theta \tag{38}$$

$$A_1(t) = Qv(t-1)\, A_0(t-1) \tag{39}$$

$$A_i(t) = v(t-1)\left(\frac{i-1}{i}\right)^{N/2} A_{i-1}(t-1) \tag{40}$$

$$\frac{1-\theta}{v(t)} = QA_0(t) + \sum_{i=1}^{t} A_i(t)\left(\frac{i}{i+1}\right)^{N/2} \tag{41}$$

For the initial conditions $A_0(0)=1$ and $A_i(0)=0$ $(i>0)$, these recursion relations have the solution

$$A_i(t) = Q \prod_{t'=t-i}^{t-1} v(t')\left(\frac{1}{i}\right)^{N/2} A_0(t-i) \tag{42}$$

The self-consistency condition is

$$\frac{1-\theta}{v(t)} = \frac{Q}{v(t)}\left[\sum_{j=1}^{t+1}\left(\frac{1}{j}\right)^{N/2} A_0(t+1-j) \prod_{t'=t+1-j}^{t} v(t')\right] \tag{43}$$

Now we separate out explicitly the term $j=t+1$; this is reasonable because $v(0)$ is larger than all the other v's by a factor proportional to $1/Q$, yielding

$$\frac{1-\theta}{v(t)} = \frac{Q\theta}{v(t)} \sum_{j=1}^{t}\left(\frac{1}{j}\right)^{N/2} \prod_{t'=t-j+1}^{t} v(t') + (1-\theta)\left(\frac{1}{t+1}\right)^{N/2} \prod_{t'=1}^{t-1} v(t') \tag{44}$$

Comparison of this result with that of the steady-state analysis (Eq. (32)) reveals that as $t\to\infty$ the last term on the rhs is just A_∞, the amplitude of the δ-function. Finally, defining $\gamma(t) = \prod_{t'=1}^{t} v(t')$, we find

$$(1-\theta) = Q\sum_{j=1}^{t}\left(\frac{1}{j}\right)^{N/2} \frac{\gamma(t)}{\gamma(t-j)} + (1-\theta)\left(\frac{1}{t+1}\right)^{N/2} \gamma(t) \tag{45}$$

Recall that at long times $v(t)$ approaches a limit $v \leqslant 1$. Therefore, for large t, $\gamma(t) = C_\gamma v^t$, where C_γ is a constant, up to corrections that vanish as $t\to\infty$, and

$$(1-\theta) = Q\sum_{j=1}^{t}\left(\frac{1}{j}\right)^{N/2} v^j + (1-\theta)\left(\frac{1}{t+1}\right)^{N/2} C_\gamma v^t \tag{46}$$

When v is strictly less than unity, the convergence is exponential, as expected from the qualitative argument given above. When $\lim_{t\to\infty} v(t) = 1$, then, since the second term on the right-hand side is nonzero as $t\to\infty$, we must have $\gamma(t) = \prod_{i=1}^{t} v(t) = C_\delta (t+1)^{N/2}$, where C_δ is a constant. This implies $v(t) \to ((t+1)/t)^{N/2} \sim 1 + \mathcal{O}(1/t)$. By calculating the correction to the sum

because of this variation in $v(t)$, we find that the amplitude of the δ-function obeys $A_t(t) - A_\infty \propto t^{-1}$ as $t \to \infty$. When $N = 2$, $v(t \to \infty)$ is exponentially close to unity, so that the decay of the amplitude in the unmutated peak up to extremely long times is governed by the logarithmic divergence of the sum for $v = 1$. Thus, in this case we expect $A_t(t)$ to decay logarithmically with t.

Figure 2 shows $A_t(t)$, the fraction of the population that is unmutated, versus time t for the parameter values $\theta = 0.2$ and $Q = 0.1$, for $N = 1, 2$, and 3. The curves were obtained by numerical iteration of the recursion relations (38)–(41), starting with the initial condition $A_0(0) = 1$ and $A_i(0) = 0$ for $i \neq 0$. We find, as expected, that $A_t(t)$ decays to zero for $N = 1$ and $N = 2$, but not for $N = 3$. The scales on the plot were chosen to emphasize the logarithmic decay when $N = 2$.

To compute the probability density $\phi(\vec{x}, t)$, we first calculate the amplitudes $A_i(t)$ using the recursion relation, and then sum the corresponding Gaussians (see Eqs. (16) and (10)) to obtain numerical values of the probability. These results are plotted in Figs. 3, 4, and 5 for the parameter-values $\theta = 0.2$ and $Q = 0.1$. For $N = 1$ (see Fig. 3), we plot $\phi(\vec{x}, t)$ against the magnitude of x for various values of the time, t. The picture shows a probability distribution that first gets narrower, but then settles down to a time-independent behavior for the largest times. Also drawn on this plot is the infinite-time probability density, derived from Eqs. (32) and (33). The density $\phi(\vec{x}, t)$ settles down to this limiting behavior most slowly near $\vec{x} = 0$, where fitnesses change slowly with $\vec{x}$.

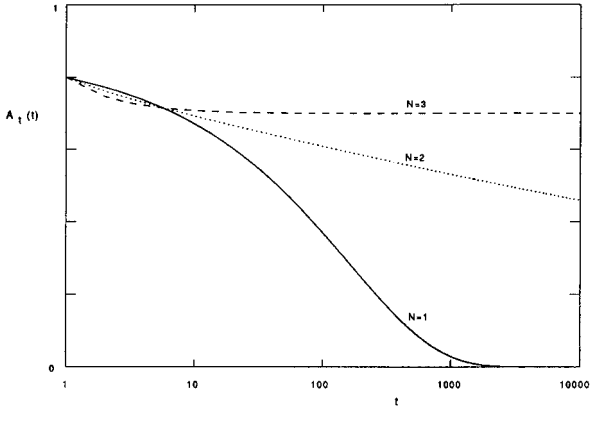

Fig. 2. Plot of the fraction of the population which is unmutated, $A_t(t)$, versus time t for the discrete-time model with parameter values $Q = 0.1$, $\theta = 0.2$, obtained by numerical iteration of Eqs. (38)–(41) starting with $A_0(0) = 1$, $A_{i \neq 0}(0) = 0$. For $N = 1$ and 2, $A_t(t)$ decays so that $A_\infty \equiv \lim_{t \to \infty} A_t(t) = 0$, while for $N = 3$, A_∞ is nonzero.

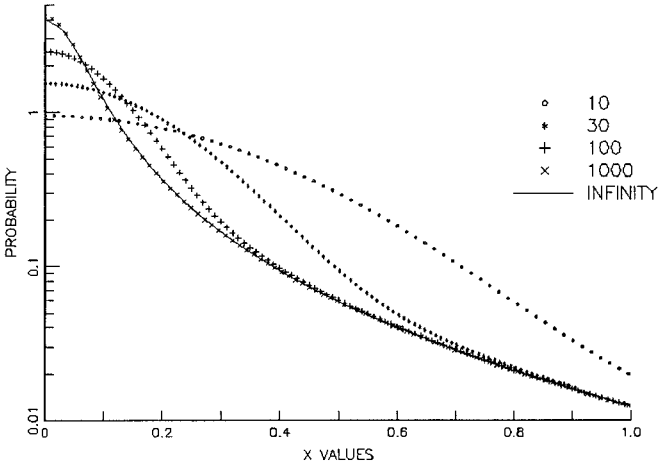

Fig. 3. Plot of probability density $\phi(\vec{x}, t)$ against the magnitude of $\vec{x}$ for $N = 1$. The numerical results for times 10, 30, 100, and 1000 are calculated by summing Gaussians with weights computed from the recursion relations. This set of curves is compared with the expected infinite-time result, calculated as a solution to Eqs. (32) and (33). For large times, the two types of calculations agree quite well.

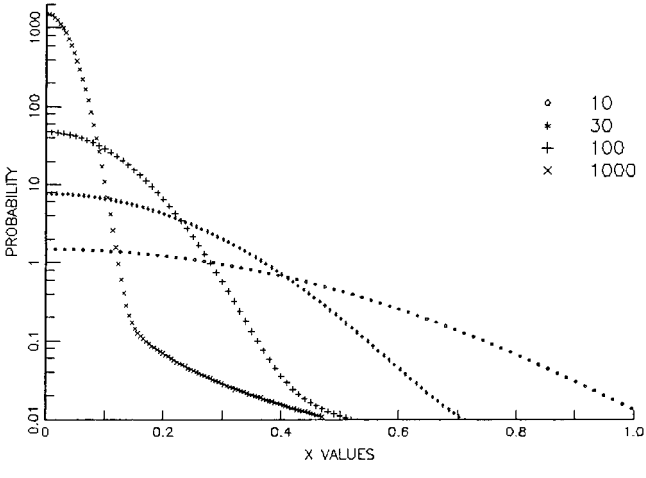

Fig. 4. Plot of density $\phi(\vec{x}, t)$ gainst the magnitude of $\vec{x}$ for $N = 3$. The numerical results are shown for times 10, 30, 100, and 1000. In contrast to the $N = 1$ case, these curves contain a central peak that continues to narrow and to grow for large t.

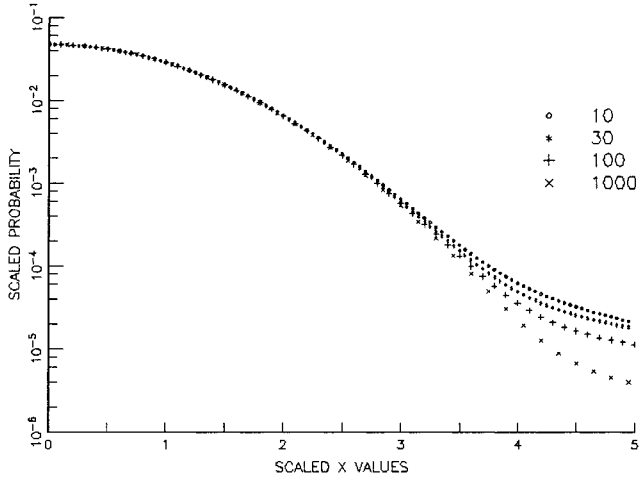

Fig. 5. Scaled plot of probability density versus the magnitude of x for $N=3$. The scaling is picked to make the central peak show a time independent behavior in the new coordinates. The abscissa is $xt^{1/2}$, so that for larger time the picture focuses upon smaller values of x. The ordinate is $\phi(\vec{x}, t)\, t^{-N/2}$, and thereby the picture focuses upon increasingly concentrated distributions for larger t. Numerical results are shown for times 10, 30, 100, and 1000. In these coordinates, the figure shows (as expected) an approach to a constant value at large times.

For $N=3$, a corresponding calculation shows a probability density that gets more and more peaked as time goes on, see Fig. 4. Notice how, for $t=1000$, the peak sticks up sharply from a more slowly varying background. To see how this peaking occurs, we plot in Fig. 5 a scaled version of Fig. 4. In this version, we plot $\phi_s(\vec{x}_s, t)$ where $\phi_s(\vec{x}, t)$ is the result of dividing $\phi(\vec{x}, t)$ by the predicted growth of the peak, proportional to $t^{N/2}$, giving us a scaled y-variable $\phi_s(\vec{x}, t) = \phi(\vec{x}, t)\, t^{-N/2}$. The scaled x-variable is $\vec{x}_s = \vec{x}t^{1/2}$. The resulting plot becomes time-independent for large times and not-too-large values of $\vec{x}_s$. In this way, we see how the peak continually gets narrower, and more and more dominates the small-x behavior.

4. CONTINUOUS-TIME EQUATIONS

We proceed by Fourier transforming Eq. (9). We define the quantity $\tilde{\phi}(\vec{k}, T) = \int d^N x\, \phi(\vec{x}, T) \exp[\, i\vec{k} \cdot \vec{x}\,]$ and obtain

$$\frac{\partial \tilde{\phi}(\vec{k}, T)}{\partial T} = \left[a(T) + \frac{1}{2V} \nabla_{\vec{k}}^2 \right] \tilde{\phi}(\vec{k}, T) - \Theta[\, 1 - f(\vec{k})\,]\, \tilde{\phi}(\vec{k}, T) \tag{47}$$

where $f(\vec{k})$ is the Fourier transform of Eq. (5),

$$f(\vec{k}) = \exp[-m^2k^2/2] \tag{48}$$

We must also specify, in addition to the initial conditions, two boundary conditions. One boundary condition is determined by the normalization requirement $\int d^N x\, \phi(\vec{x}, T) = 1$, which yields

$$\tilde{\phi}(\vec{0}, T) = 1 \tag{49}$$

The second boundary condition is that $\tilde{\phi}(\vec{k}, T) \leqslant 1$ for all $\vec{k}$, which in particular implies that it cannot diverge as $\vec{k} \to \infty$. This follows from combining the normalization condition with the non-negativity requirement $\phi(\vec{x}, T) \geqslant 0$:

$$\tilde{\phi}(\vec{k}, T) = \int d^N x\, e^{i\vec{k}\cdot\vec{x}} \phi(\vec{x}, T) \leqslant \int d^N x\, \phi(\vec{x}, T) = 1 \tag{50}$$

We now solve Eq. (47) in the limit of long times. Since as $T \to \infty$ both $\tilde{\phi}$ and a become independent of time, we must solve

$$\{ -\tfrac{1}{2}\nabla_{\vec{k}}^2 - (\Theta V)\, f(\vec{k})\}\ \psi(\vec{k}) = E\psi(\vec{k}) \tag{51}$$

where $\psi(\vec{k}) \equiv \lim_{T\to\infty} \tilde{\phi}(\vec{k}, T)$ and the eigenvalue $E = V(\lim_{T\to\infty} a(T) - \Theta)$. Equation (51) is just the time-independent Schrödinger equation. The parameter N, the number of characters affected by each mutation, here is interpreted as the spatial dimensionality. As shown in Bürger and Bomze[10] and references therein, and as discussed here in Section 5, the long-time solution to the population biology model is given by the lowest energy eigenstate of this Schrödinger equation.

The kinetic energy term $-\frac{1}{2}\nabla_{\vec{k}}^2$ in Eq. (51) comes from selection. This term, which in $\vec{x}$-space causes the distribution to become progressively narrower, acts to make it progressively wider in $\vec{k}$-space. The potential energy term $-\Theta V f(\vec{k})$ comes from mutation. This potential is constant at large k, and has an attractive piece at small k. If we assume that the potential has a characteristic scale in k, so that $f(\vec{k}) = F(m\vec{k})$ (clearly, the fitness function considered here is of this form), then we can define $\vec{z} = m\vec{k}$ and write the Schrödinger equation in dimensionless form:

$$\left\{ -\frac{1}{2}\nabla_{\vec{z}}^2 - \left(\frac{\Theta V}{m^2}\right) F(\vec{z}) \right\} \psi(\vec{z}) = \tilde{E}\psi(\vec{z}) \tag{52}$$

where $\tilde{E} = E/m^2$. This equation makes it clear that the nature of the behavior depends only on the single parameter $(\Theta V/m^2)$.

If only the selection (kinetic energy) term were present, the solutions to Eq. (51) consistent with the normalization condition $\phi(\vec{k}) = 1$ would be

$$\psi(\vec{z}) = \exp[\, i\vec{p} \cdot \vec{z}\,] \tag{53}$$

where $\vec{p} = \hat{z}\sqrt{2\tilde{E}}$, and $\hat{z}$ is a unit vector along $\vec{z}$. To be consistent with the boundary conditions, we require $\tilde{E} \geqslant 0$. The ground state eigenstate thus has $\tilde{E} = 0$, so that $\psi(\vec{z}) = 1$. Fourier transforming this result, we see that in the absence of mutation, at $T = \infty$ the entire population has the same genotype.

Now we consider the effects of adding the potential arising from the mutation term. This potential is attractive, so that the question we must address is whether the ground state eigenstate remains extended ($\psi(\vec{z}) \to$ constant as $|\vec{k}| \to \infty$) or whether it results in a bound state with $\tilde{E} < 0$, which has $\psi(\vec{z}) \to 0$ exponentially in k as $\vec{k} \to \infty$. In the latter case, the real-space distribution $\phi(\vec{x})$ is smooth as $\vec{x} \to 0$. Therefore, if there is a bound state, then only an infinitesimal fraction of the population is unmutated, whereas if there are no bound states, then a finite fraction of the population has a unique genotype.

It has been proven for the Schrödinger equation that any attractive potential, no matter how small, will lead to a bound state for $N \leqslant 2$.[9] Thus, the population distribution in real space as $T \to \infty$, $\phi(\vec{x}, \infty)$, must be smooth for $N \leqslant 2$. When $N > 2$, bound states appear only if the potential is large enough ($\Theta V_s/m^2 > C_N$, where C_N is of order unity).[8] Therefore, in this case, if the potential is small (weak mutation), then the ground state remains extended: $\psi(\vec{k}) \to$ constant as $\vec{k} \to \infty$, and the real space distribution $\phi(\vec{x})$ has a δ-function piece. If the potential is large (strong mutation), then there is a bound state, and the distribution $\phi(\vec{x})$ is not singular at small $\vec{x}$.

If all the states are extended, then as $|\vec{k}| \to \infty$, the lowest energy state obeys $\nabla_{\vec{k}}^2 \tilde{\phi}(\vec{k}) = 0$ and has the form $\tilde{\phi}(\vec{k}) \to A + Bk^{-(N-2)}$, where A and B are coefficients determined by matching to the solution in the small $|\vec{k}|$ region. For almost all potentials, both A and B are nonzero; Fourier transforming the term proportional to B yields

$$\text{B term} \propto \int d\Omega \int_0^\infty dk\, k^{N-1} e^{ikx\cos(\theta)} k^{-(N-2)} \propto x^{-2} \tag{54}$$

(Here, $\int d\Omega$ is the integral over angles.) Thus, just as in our discrete-time solution, we find associated with a δ-function at $\vec{x} = 0$ a power-law divergence at small x: $\phi(\vec{x}) \propto x^{-2}$.

In the regime where the ground state is bound, there is a finite energy gap between the ground state and the excited states, which implies that the long-time limit is approached exponentially in time. When there is no bound state, the energy spectrum of the Schrödinger equation is continuous, which leads to power-law convergence to the long-time limit. These results are consistent with those from the discrete-time model, and are discussed further in Section 5.

5. RELATION TO TIME-DEPENDENT QUANTUM MECHANICS

In this section we relate the discrete-time model to the quantum-mechanical problem of a particle in the presence of a potential that is periodic in time. This correspondence is very useful because it enables us to show that the probability distribution in the limit of infinite time is the same for all initial conditions[9] and to argue that the time-dependence of the approach to this limiting distribution is the same for different initial conditions and for different parameter regimes. For the continuous-time model, which corresponds to the case of a time-independent potential, the result that the longtime behavior of Eq. (47) is given by the lowest energy eigenstate of the associated Schrödinger Eq. (51) has been proved by Bürger and collaborators [10, and references therein].

We write the discrete-time model for general τ as a problem of the quantum mechanics of a particle in the presence of a time-dependent potential consisting of δ-function "kicks" applied at discrete times $T = t\tau$, where t is an integer. To do this, we define $\Phi_-(\vec{x}, t\tau) = \lim_{\varepsilon \to 0} \Phi(\vec{x}, t\tau - \varepsilon/2)$ and $\Phi_+(\vec{x}, t\tau) = \lim_{\varepsilon \to 0} \Phi(\vec{x}, t\tau + \varepsilon/2)$, identify $\Phi_+(\vec{x}, t\tau) = \phi(\vec{x}, t)$, and write Eq. (2) as

$$\Phi_-(\vec{x}, t\tau) = \frac{\Phi_+(\vec{x}, (t-1)\,\tau) \exp[-(\vec{x}^2/2V_s)]}{\int d^N x\, \Phi(\vec{x}, (t-1)\,\tau) \exp[-(\vec{x}^2/2V_s)]} \tag{55}$$

$$\Phi_+(\vec{x}, t\tau) = \Phi_-(\vec{x}, t\tau) + \Theta\tau \int d^N y\, f(\vec{x} - \vec{y})[\Phi_-(\vec{y}, t\tau) - \Phi_-(\vec{x}, t\tau)] \tag{56}$$

Here, the continuous time variable $T = t\tau$, and $\Theta = \theta/\tau$ is the mutation rate per unit time.

First consider Eq. (55). Note that the differential equation

$$\frac{\partial \Phi(\vec{x}, T)}{\partial T} = \left[a(T) - \frac{\vec{x}^2}{2V} \right] \Phi(\vec{x}, T), \qquad (t-1)\,\tau + \frac{\varepsilon}{2} < T < t\tau + \frac{\varepsilon}{2} \tag{57}$$

[9] The uniqueness of the long time follows from general properties of linear operators[24] and has been pointed out for this problem by Bürger and Bomze.[10]

has the solution

$$\Phi(\vec{x}, T+\tau) = \exp\left[\int_T^{T+\tau} ds\, a(s) \right] \exp\left[-\frac{\vec{x}^2}{2V} \tau \right] \Phi(\vec{x}, T) \tag{58}$$

Thus, we identify $V_s = V/\tau$, and choose $a(T)$ so that[10]

$$\exp\left[-\int_T^{T+\tau} ds\, a(s) \right] = \int d^N x\, \Phi(\vec{x}, T) \exp\left[-\frac{x^2}{2V_s} \right] \tag{59}$$

Recalling Eq. (55), we see that this differential equation transforms $\Phi_+(\vec{x}, T)$ into $\Phi_-(\vec{x}, T+\tau)$.

We now Fourier transform Eq. (56), yielding

$$\tilde{\Phi}_+(\vec{k}, t\tau) = \exp[-W(\vec{k})]\, \tilde{\Phi}_-(\vec{k}, t\tau) \tag{60}$$

with $\tilde{\Phi}_\pm(\vec{k}, t\tau) = \int d^N x \exp[i\vec{k}\cdot\vec{x}]\, \Phi_\pm(\vec{x}, t\tau)$, $\tilde{f}(\vec{k}) = \int d^N x \exp[i\vec{k}\cdot\vec{x}]\, f(\vec{x})$, and

$$W(\vec{k}) = -\ln[1 - \Theta\tau(1 - \tilde{f}(\vec{k}))] \tag{61}$$

Equation (60) is the solution to the differential equation[11]

$$-\frac{\partial \tilde{\Phi}(\vec{k}, T)}{\partial T} = W(\vec{k})\, \delta(T - t\tau)\, \tilde{\Phi}(\vec{k}, T), \quad t\tau - \frac{\varepsilon}{2} < T < t\tau + \frac{\varepsilon}{2} \tag{62}$$

Letting $\varepsilon \to 0$, Fourier transforming Eq. (57), and using Eq. (62), we obtain

$$-\frac{\partial \tilde{\Phi}(\vec{k}, T)}{\partial T} = \left[-a(T) - \frac{1}{2V} \nabla_k^2 \right] \tilde{\Phi}(\vec{k}, T) + W(\vec{k})\, \delta(T - t\tau)\, \tilde{\Phi}(\vec{k}, T), \qquad (t-1)\,\tau + \frac{\varepsilon}{2} < T < t\tau + \frac{\varepsilon}{2} \tag{63}$$

[10] As seen above in Eq. (9), the choice $a(T) = \int d^N x\, x^2/2V\, \Phi(\vec{x}, T)$ ensures that $\int d^N x\, \Phi(\vec{x}, T) = 1$ for all T.

[11] This step follows because Eq. (62) can be rewritten as

$$-\int_{\ln \tilde{\Phi}(\vec{k}, t\tau - \varepsilon)}^{\ln \tilde{\Phi}(\vec{k}, t\tau + \varepsilon)} d \ln \tilde{\Phi}(\vec{k}, T) = W(\vec{k}) \int_{t\tau - \varepsilon}^{t\tau + \varepsilon} dT\, \delta(T - t\tau)$$

Extension to $T \geqslant 0$ requires summing the last term over $t = 0, 1, 2, \ldots$, and we can further extend this differential equation to $-\infty < T < \infty$ by defining $\tilde{\Phi}(\vec{k}, T)$ for $T < 0$ by solving it backwards in time, starting at $T = 0$. Therefore, the Fourier transform of the discrete-time model Eq. (2) can be written as the differential equation[12]

$$-\frac{\partial \tilde{\Phi}(\vec{k}, T)}{\partial T} = \left[-a(T) - \frac{1}{2V} \nabla_{\vec{k}}^2 \right] \tilde{\Phi}(\vec{k}, T) + \sum_{t=-\infty}^{\infty} \delta(T - t\tau)\, W(\vec{k})\, \tilde{\Phi}(\vec{k}, T) \tag{64}$$

In dynamical systems theory, equations with a sum of time-delta functions are said to be "kicked".[25]

We now write

$$\tilde{\Phi}(\vec{k}, T) = \exp\left[\int_0^T ds\, a(s) \right] \xi(\vec{k}, T) \tag{65}$$

Because $\tilde{\Phi}(\vec{k}, T)$ satisfies Eq. (64), $\xi(\vec{k}, T)$ must obey

$$-\frac{\partial \xi(\vec{k}, T)}{\partial T} = -\frac{1}{2V} \nabla_{\vec{k}}^2 \xi(\vec{k}, T) + \sum_n \delta(T - t\tau)\, W(\vec{k})\, \xi(\vec{k}, T) \tag{66}$$

Equation (66) describes an imaginary-time quantum mechanics problem in which the Hamiltonian is periodic in time.[13] This time-periodicity implies that all the solutions of Eq. (66) can be written in the form[27–29]

$$\xi(\vec{k}, T) = \sum_m A_m \exp[-\lambda_m T]\, u_m(\vec{k}, T) \tag{67}$$

where each A_m is independent of time and the quasi-eigenstates $u_m(\vec{k}, T)$ are orthonormal and time-periodic, $u_m(\vec{k}, T + \tau) = u_m(\vec{k}, T)$. Therefore,

$$\tilde{\Phi}(\vec{k}, T) = \exp\left[\int_0^T ds\, a(s) \right] \sum_m A_m \exp[-\lambda_m T]\, u_m(\vec{k}, T) \tag{68}$$

We denote the smallest λ_m as λ_0. As T increases, the relative weight in this quasi-eigenstate grows exponentially. Therefore, unless the initial

[12] When τ, the interval between kicks, goes to zero, $W(\vec{k}) \to \Theta\tau(1 - f(\vec{k}))$ and the sum of delta-functions is replaced by its time average, so that Eq. (64) reduces to Eq. (47), as expected from the discussion in Section 2 above.

[13] Partial differential equations of this type have been studied in the context of stochastic resonance[26] as well as periodically driven quantum-mechanical system.[27–31]

condition is such that the amplitude A_0 is exactly zero, at long times this quasi-eigenstate dominates, and as $T \to \infty$, $\Phi(\vec{k}, T) \propto u_0(\vec{k}, T)$. Sturm–Liouville theory guarantees that the ground state is nodeless[32] and therefore not orthogonal to a nonnegative initial condition. Thus, all initial conditions yield the same behavior in the limit of long times.[14]

Now we discuss the approach to the long-time limit. At any finite time T, the ratio of the weights of any two quasi-eigenstates i_1 and i_2 is proportional to $\exp[-(\lambda_{i_1} - \lambda_{i_2})\, T]$. If there is a nonzero energy gap between the lowest and second-lowest energy quasi-eigenstates (no δ-function in $\phi(\vec{x})$), then the approach to the long-time limit is exponential in time. If there is no bound state, then all the states are extended and the spectrum is continuous; because the potential is of finite range and the wavefunctions are extended, the λ_i take the form $\lambda_i \propto q_i^2/2$, where q_i has the dimension of a wavevector in $\vec{k}$ space[8] and hence of a distance in the original $\vec{x}$ space. Therefore,

$$\xi(\vec{k}, t) = \sum_m A_m e^{-q_m^2 T/2} \tag{69}$$

The relative weight of the eigenstate m remains large until $q_m \gtrsim \sqrt{1/T}$. Thus, the δ-function at infinite times emerges from a peak that is narrowing, having a width proportional to $\sqrt{1/\text{time}}$.

This result for the time-dependence of the approach to the long time distribution is completely consistent with the explicit calculations presented in Section 3. Because it does not depend on the magnitude of V_s/m^2, it indicates that the time-dependence of the approach to the long-time limit is not sensitive to the approximation $V_s/m^2 \ll 1$ made in that section.

6. COMPARISON WITH SEQUENCE DATABASE STATISTICS

Here we attempt to exploit the public availability of extensive DNA sequence databases and tools for comparing them to assess empirically the relevance of the pleiotropy model. These databases are comprehensive compendia of observed DNA sequence variation, for submission of sequence data to a database is requirement for publication in all biological journals. The aim is to test a qualitative prediction of the model; that genes with a high degree of pleiotropy should have a narrower distribution of alleles than those that affect only one trait. This trend, which is consistent with the mathematical results of the previous sections, is easy to understand: if

[14] This result requires the quasi-ground state to be nondegenerate, which can be demonstrated explicitly when $\Theta\tau$ is small, and hence should be true generically.

a gene of high fitness is pleiotropic, then because each mutation affects several characters independently, the chances are high that a given mutation leads to a large fitness decrease.

Our first test for possible correlation between degree of pleiotropy and the probability distribution describing genetic variation is simply to count the number of naturally occurring variants (alleles) of various *Drosophila melanogaster* genes using FlyBase.[33] We choose to include only naturally occurring *D. Melanogaster* alleles, including spontaneous mutations arising in laboratory stocks. Table I shows data for nine genes: *brown*, *cinnabar*, *ecdysone receptor*, *engrailed*, *fork head*, *hairless*, *notch*, *vestigial*, and *white*. The number of variants is tabulated together with the length of the primary transcript and the approximate length of the genomic region encompassing the gene and its flanking regulatory regions. The number of variants per gene is normalized both by the transcript length and by the genomic length. In both cases, the values range over approximately three orders of magnitude. We estimate the degree of pleiotropy of each of the listed genes according to the number of tissues/structures and developmental stages in which the genes are expressed. This ordering is admittedly arbitrary to some extent, but few would argue with the judgment that *fork head* (encoding a transcription factor essential for development of the midgut) and *engrailed* (encoding a homeotic gene establishing segmentation) are more pleiotropic than the eye color mutants *cinnabar*, *white*, and *brown*. Table I reveals a

Table I. Number of Naturally-Occurring Mutations for Different *Drosophila Genes*[a]

Gene	PR	N_V	L_T	L_G	N_V/L_T	N_V/L_G
fork head	1	1	4.0	13–60	2.5×10^{-4}	$(1.7–7.7)\times10^{-5}$
engrailed	2	19	3.6	200	5.3×10^{-4}	9.5×10^{-5}
ecdysone receptor	3	13	5.5	35	2.4×10^{-3}	3.7×10^{-4}
notch	4	70	15	80–150	4.7×10^{-3}	$(4.7–8.8)\times10^{-4}$
vestigial	5	55	3.8	46	0.014	1.2×10^{-3}
hairless	6	17	6.0	8	2.0×10^{-3}	2.1×10^{-3}
cinnabar	7	22	2.5	15	8.8×10^{-3}	1.5×10^{-3}
white	8	247	6.0	48–200	0.041	$(1.2–5.1)\times10^{-3}$
brown	9	61	3.0	140	0.020	4.4×10^{-4}

[a] PR is pleiotropy rank, estimated using the number of tissues/structures and developmental stages in which the genes are expressed. N_V is the number of naturally-occurring variants of the gene, L_T is the transcript length, and L_G is the genomic length. All lengths are given in kilobases. Transcript and genomic lengths are taken from the full-format gene reports in FlyBase. For genes with multiple transcripts, we report the length of the longest described transcript. When the full gene report does not include the genomic length, this is estimated from the FlyBase molecular map of the gene.

Table II. Genes Examined Using BLAST 2 Sequence Similarity Search Tool[a]

Gene	Length	Number of matches
cytochrome P-1-450	2565	81
dystrophin	13957	216
epidermal growth factor receptor	2660	57
HOX A1 (human)	2595	380
huntingtin	10348	108
idothyronine deiodinase	2222	159
rhodopsin	6953	208
alpha-tubulin	1596	488

[a] Lengths of genes are given in units of the number of bases.

clear inverse correlation between estimated pleiotropy and normalized number of variants, consistent with the prediction of Waxman and Peck's model.

A limitation of this approach is that other factors besides degree of pleiotropy could lead to the observed differences in the number of variants. For example, a gene that controls eye color may well have a much smaller overall effect on fitness than a gene that controls a crucial developmental function. In the population genetics model, differences of this type are reflected by different values of the fitness parameter V_s. Within the model, differences in overall variability caused by changing V_s can be distinguished from those caused by changing the degree of pleiotropy N because they yield very different functional forms for the distribution function $\phi(\vec{x})$. However, counting alleles yields no information about the functional form of the distribution function and thus cannot be used to distinguish between the mechanisms.

Our second strategy for relating degree of pleiotropy to statistics of archived DNA sequences aims to obtain information about the form of the probability distribution describing the variation for a set of genes. It assumes that two genetic variants whose sequences are highly similar in sequence space code for genes that are close together in fitness space. We use the BLAST 2 sequence similarity search tool[34] to search the GENBANK and EMBL DNA sequence databases[15] against the 8 gene sequences listed in Table II. These databases contain sequences of genes from many organisms, though a preponderance are from humans. Given a target sequence, BLAST 2 generates a list of matching sequences along

[15] These databases and the BLAST 2 search tool are maintained by the National Center for Biotechnology Information (http://www.ncbi.nlm.nih.gov).

with scores which measure the degree of similarity. Because the maximum possible score for a given sequence is larger for longer sequences, we avoided examining either very short or very long genes, but even so, the lengths of the sequences examined varied by nearly a factor of ten. As shown in Table II, the number of matches obtained also varies by roughly an order of magnitude for the genes in the table, and is not obviously correlated with the length of the gene.

High BLAST 2 scores correspond to sequences which match the search sequence the most closely. The scores range from roughly 10^4 (a sequence matching itself) to the default cutoff of 40. We assume that this score is inversely correlated with the evolutionary distance between the input gene and the retrieved sequences, so that a gene with a high fraction of matches with low scores has a broader probability distribution $\phi(x)$ than a gene with more matches with high scores. Thus, if all the matches tend to be very good, the gene is considered to be less variable than one with many poor matches.

Figure 6 is a histogram of the fraction of the matches in a given score range obtained versus the inverse of the score for several genes. There is significant variability between these genes; for example, *cytochrome P-1-450*, which encodes a general purpose antioxidant expressed in many tissues and is thus plausibly highly pleiotropic, has relatively many very good matches

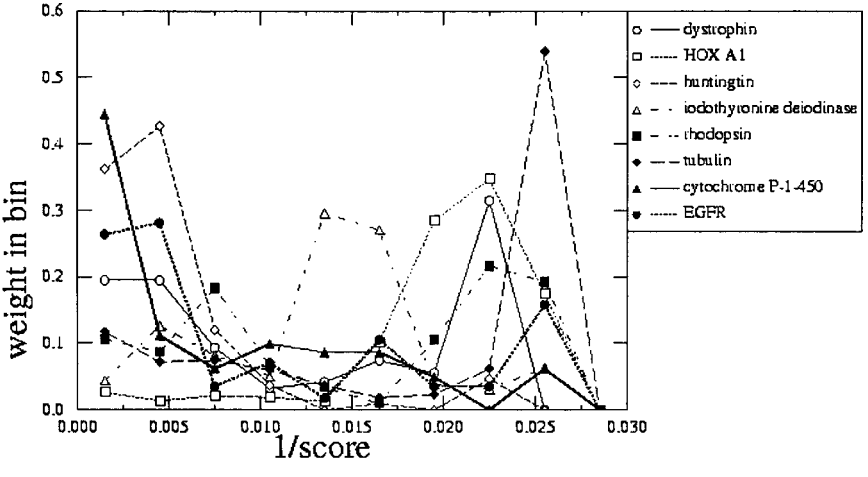

Fig. 6. Histogram plot of the fraction of matches with inverse scores in a given range, as a function of inverse score, for several genes with differing degrees of pleiotropy. Significant differences in the statistics of these matches are observed; for example, *cytochrome P-1-450* has a very large percentage of matches with high scores, whereas the matches for *rhodopsin* tend to have low scores.

compared to *rhodopsin*, which is quite specific, encoding a protein essential for vision. However, *HOX A1* has many poor matches, where it might be expected to be pleiotropic, since it is essential to many developmental functions. Thus, though interesting variations in the statistics of DNA sequences for different genes are found, we have not been able to demonstrate convincingly a correlation between degree of pleiotropy and these statistical variations.

The behavior of *HOX A1* can be explained by a process of repeated duplication and divergence over the course of evolution. Many of the retrieved genes have assumed distinct but related developmental functions. The BLAST 2 analysis presented here does not account for a shift in score distributions arising from divergence among members of gene families and superfamilies.

There are several additional difficulties and ambiguities inherent in comparisons between genome sequence data and population genetics models of the type considered here. First, the BLAST scores are based on the quality of sequence alignments, whereas the population genetics models define distances in terms of fitness, which is a phenotypic quantity. Fitness distance and evolutionary sequence distance may be quite different; for example, point mutations at different locations may range in effect from unobservable to lethal. However, even if there is not a single definite relationship between sequence distance and fitness distance, so long as these quantities are positively correlated with each other, meaningful results may be obtained, if all the genes are all subject to similar uncertainties. Another limitation of our analysis is that we consider the fitnesses of alleles as fixed. This is not necessarily the case, as many gene products function as components of multi-protein structures or pathways.[35, 36] This dependency of the fitness value of a specific sequence variant on the remainder of the genotype is manifested explicitly by the existence of epistatic interactions (e.g., refs. 37 and 38).

In addition, one may worry that the population genetics model we have used involves a continuously varying phenotype, whereas sequence variations are discrete. However, because genes are thousands of base pairs long, there is still a huge range of variability, and the use of a continuum model is therefore reasonable.

Finally, pleiotropy itself is defined in terms of phenotypic fitness, and it is not clear how to create an independent measure that enables one to assign objectively the degree of pleiotropy N to a given gene. This problem is illustrated by the case of *alpha-tubulin*, which codes for a protein vital to the formation of microtubules. On the one hand, microtubules are expressed in many different tissues, but on the other hand, all its functions arise from similar structural properties. Thus, one could categorize *alpha-tubulin* as

either highly pleiotropic (since it is expressed in many tissues) or as non-pleiotropic (since the function is similar everywhere where it is expressed). This is a serious fundamental difficulty that must be overcome if one is to make meaningful quantitative statistical analysis of the possible correlations between the sequence statistics and the degree of pleiotropy.

7. DISCUSSION

In this paper we have analyzed some variants of a population biology model incorporating selection, mutation, and pleiotropy. We have focused on understanding the circumstances under which at long times a nonzero fraction of the population has a unique genotype, and on characterizing the time dependence of the population distribution in this regime. We have analyzed the discrete-time model of Waxman and Peck[1] as well as an associated continuous-time model. We find:

1. In both the discrete and continuous-time models, a unique genotype can emerge only when N, the number of characters affected by each gene, is greater than two, a result in agreement with Waxman and Peck.[1] For any $N > 2$, the infinite-time population distribution $\phi(\vec{x}, \infty)$ contains a δ-function contribution when the mutation rate is nonzero but small, but not when the mutation rate is large enough.

2. In the regime where $\phi(\vec{x}, \infty)$ has a δ-function contribution, this δ-function is accompanied by a $1/x^2$ singularity at small $\vec{x}$.

3. The δ-function peak emerges as the limit of a peak that continually becomes higher and narrower. Thus, in this regime the convergence to the $t \to \infty$ limit is a power law.

4. In the regime when $\phi(\vec{x}, \infty)$ is smooth, the convergence to this distribution is exponential in time.

5. The continuous- and discrete-time models exhibit qualitatively identical behavior, but there are quantitative differences between them.

Our analysis of the discrete-time equations relies on the use of Gaussian functions for both the fitness and mutation terms. Our analysis of the continuous-time model assumes a Gaussian fitness function, but it does not assume that the mutation term is Gaussian; it applies for a large class of different mutation terms. This is because in the quantum mechanics problem the kinetic energy is always a second derivative, but similar results are obtained for any short-ranged attractive potential.

Other evolutionary forces or genetic assumptions, such as antagonistic pleiotropy, more complex fitness landscapes, and discreteness of alleles,

change qualitatively the nature of the equations describing the system. It will be interesting to see whether the "condensation" phenomenon investigated here is robust when these effects are taken into account.

We also present in Section 6 an assessment of whether the degree of pleiotropy of selected genes result in systematic trends in their mutation spectra documented in online databases. These investigations are inconclusive. At present, available databases do not allow unbiased sampling of the sequence variation present in a natural population. Such a database is likely to emerge as the Human Genome Project's present initiative to sample the extent of sequence variation in a wide array of genes in the American population progresses.

ACKNOWLEDGMENTS

We are grateful for support by the National Science Foundation, Grant DMR 96-26119 (SNC), and the Office of Naval Research, Grant N00014-96-1-0127 (LPK). We thank Dr. J. J. Sohn and Prof. T. Nagylaki for useful suggestions.

REFERENCES

1. D. Waxman and J. R. Peck, Pleiotropy and the preservation of perfection, *Science* **279**:1210–1213 (1998).
2. M. Kimura, A stochastic model concerning the maintenance of genetic variability in quantitative characters, *Proc. Nat. Acad. Sci. USA* **54**:731–736 (1965).
3. R. Lande, The genetic covariance between characters maintained by pleiotropic mutation, *Genetics* **94**:203–215 (1980).
4. M. Turelli, Effects of pleiotropy on predictions concerning mutation-selection balance for polygenic traits, *Genetics* **111**:165–195 (1985).
5. G. P. Wagner, Multivariate mutation-selection balance with constrained pleiotropic effects, *Genetics* **122**:223–234 (1989).
6. S. Gavrilets and G. de Jong, Pleiotropic models of polygenic variation, stabilizing selection, and epistasis, *Genetics* **134**:609–625 (1993).
7. S. R. de Groot, G. Y. Hooyman, and C. A. ten Seldam, On the Bose–Einstein condensation, *Proc. Roy. Soc. A* **203**:266–286 (1950).
8. L. I. Schiff, *Quantum Mechanics*, 3rd ed. (McGraw–Hill, New York, 1968).
9. B. Simon, The bound state of weakly coupled Schrödinger operators in one and two dimensions, *Ann. of Phys.* **97**:279–288 (1976).
10. R. Burger and I. M. Bomze, Stationary distributions under mutation-selection balance: Structure and properties, *Adv. Appl. Prob.* **28**:227–251 (1996).
11. J. F. Crow and M. Kimura, *An Introduction to Population Genetics Theory* (Harper and Row, New York, 1970).
12. M. Turelli, Heritable genetic variation via mutation-selection balance: Lerch's zeta meets the abdominal bristle, *Theoret. Pop. Biol.* **25**:138–193 (1984).

13. J. F. C. Kingman, A simple model for the balance between selection and mutation, *J. Appl. Prob.* **15**:1–12 (1978).
14. W. H. Barton and M. Turelli, Evolutionary quantitative genetics: how little do we know? *Annu. Rev. Genet.* **23**:337–370 (1989).
15. R. Bürger, On the maintenance of genetic variation: global analysis of Kimura's continuum-of-alleles model, *J. Math. Biol.* **2**:341–351 (1986).
16. R. Bürger, Mutation-selection balance and continuum-of-alleles models, *Mathematical Biosciences* **91**:67–83 (1986).
17. R. Bürger and J. Hofbauer, Mutation load and mutation-selection-balance in quantitative genetic traits, *J. Math. Biol.* **32**:193–218 (1994).
18. G. F. Carrier, M. Krook, and C. E. Pearson, *Functions of a Complex Variable* (Hod books, Ithaca, New York, 1983), p. 192.
19. M. Abramowitz and C. A. Stegun (eds.), Riemann zeta function and other sums of reciprocal powers, in *Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables*, Chapter 23.2, 9th printing (New York, Dover, 1972), pp. 807–808.
20. L. Lewin, *Polylogarithms and Associated Functions* (North-Holland, New York, 1981).
21. D. Mermin and H. Wagner, Absence of ferromagnetism or antiferromagnetism in one- or two-dimensional isotropic Heisenberg models, *Phys. Rev. Lett.* **17**:1133–1136 (1966).
22. P. C. Hohenberg, Existence of long-range order in one and two dimensions, *Phys. Rev.* **158**:383–386 (1967).
23. J. Fröhlich and C. Pfister, On the absence of spontaneous symmetry breaking and of crystalline order in two-dimensional systems, *Commun. Math. Phys.* **81**:277–298 (1981).
24. M. Reed and B. Simon, *Methods of Modern Mathematical Physics, Vol. 4. Analysis of Operators* (Academic Press, New York, 1972).
25. A. J. Lichtenberg and M. A. Liberman, *Regular and Chaotic Dynamics*, Springer Verlag Applied Mathematical Sciences series, No. 38 (Springer Verlag, New York, 1992), p. 219.
26. P. Jung and P. Hänggi, Stochastic nonlinear dynamics modulated by external periodic forces, *Europhys. Lett.* **8**:505–510 (1989).
27. Ya. B. Zel'dovich, The quasienergy of a quantum-mechanical system subjected to a periodic action, *Sov. Phys. JETP* **24**:1006–1008 (1967).
28. V. I. Ritus, Shift and splitting of atomic energy levels by the field of an electromagnetic wave, *Sov. Phys. JETP* **24**:1041–1044 (1967).
29. Ya. B. Zel'dovich, Scattering and emission of a quantum system in a strong electromagnetic wave, *Sov. Phys.-Usp.* **16**:427–433 (1973).
30. D. R. Grempel, S. Fishman, and R. E. Prange, Localization in an incommensurate potential: An exactly solvable model, *Phys. Rev. Lett.* **49**:833-836 (1982).
31. D. R. Grempel, R. E. Prange, and S. Fishman, Quantum dynamics of a nonintegrable system, *Phys. Rev. A* **29**:1639–1647 (1984).
32. P. M. Morse and H. Feshbach, *Methods of Theoretical Physics, Part I* (McGraw–Hill, New York, 1953).
33. The FlyBase Consortium, FlyBase—A *Drosophila* Database, *Nucleic Acids Research* **26**:85–88 (1988); http://flybase.bio.indiana.edu/.
34. S. F. Altschul, T. L. Madden, A. A. Schäffer, J. Zhang, Z. Zhang, W. Miller, and D. J. Lipman, Gapped BLAST and PSI-BLAST: A new generation of protein database search programs. *Nucleic Acids Res.* **25**:3389–3402 (1997); http://www.ncbi.nlm.nih.gov/BLAST/.
35. R. C. Lewontin, *The Genetic Basis of Evolutionary Change* (Columbia University Press, New York, 1974).
36. G. A. Dover and R. B. Flavell, Molecular coevolution: DNA divergence and the maintenance of function, *Cell* **38**:622–623 (1984).

37. R. J. Fijneman, S. S. de Vries, R. C. Jansen, and P. Demant, Complex interactions of new quantitative trait loci, Sluc1, Sluc2, Sluc3, and Sluc4, that influence the susceptibility to lung cancer in the mouse, *Nat. Genet.* **14**:465–467 (1996).
38. T. van Wezel, A. P. Stassen, C. J. Moen, A. A. Hart, M. A. van der Valk, and P. Demant, Gene interaction and single gene effects in colon tumour susceptibility in mice, *Nat. Genet.* **14**:468–470 (1996).