

**Small world effects in evolution**

Franco Bagnoli\*

*Dipartimento di Matematica Applicata, Università di Firenze, Via S. Marta, 3 I-50139 Firenze, Italy  
and INFN, INFN, Sezione di Firenze, Largo E. Fermi, 2 I-50125 Firenze, Italy*

Michele Bezzi†

*SISSA Programme in Neuroscience, Via Beirut 2-4 I-34103, Trieste, Italy  
and INFN, Sezione di Firenze, Largo E. Fermi, 2 I-50125 Firenze, Italy*

(Received 31 July 2000; revised manuscript received 9 February 2001; published 26 July 2001)

For asexual organisms point mutations correspond to local displacements in the genotypic space, while other genotypic rearrangements represent long-range jumps. We investigate the spreading properties of an initially homogeneous population in a flat fitness landscape, and the equilibrium properties on a smooth fitness landscape. We show that a small-world effect is present: even a small fraction of quenched long-range jumps makes the results indistinguishable from those obtained by assuming all mutations equiprobable. Moreover, we find that the equilibrium distribution is a Boltzmann one, in which the fitness plays the role of an energy, and mutations that of a temperature.

DOI: 10.1103/PhysRevE.64.021914

PACS number(s): 87.10.+e, 05.20.-y, 05.45.-a

**I. INTRODUCTION**

Darwinian evolution on asexual organisms acts by two mechanisms: *mutations*, which increase the genetic diversity, and *selection*, which fixes and reduces this diversity. We classify mutations into point mutations, corresponding to local displacements on the genotypic space (defined more accurately in the following), and other types of mutation or rearrangement, which imply larger jumps. Mutations are quite rare, so we can safely assume that for each generation at most only one mutation occurs. From a biological point of view, this discussion applies to unicellular asexual organisms, for which there is no distinction between somatic and germ cells. Moreover, we do not consider the possibility of recombination (exchange of genetic material). For nonrecombinant (asexual) organisms, the combined effects of reproduction and mutations correspond to a random walk on the genotypic space. Even sexual (recombinant) populations do transmit asexually some part of their genotype, such as mitochondrial or y-chromosome DNA, to which the following analysis applies. Furthermore, we assume a constant environment.

We schematize the effects of selection by the selective reduction of the survival probability. Selection results from several constraints and can affect the frequency of certain genotypes, or even eliminate them from the population. We classify the components of selection into static and dynamic ones. In the first class we put the constraints that are independent of the population distribution, such as the functionality of a certain protein or the tuning of a metabolic path. The static part of the selection is equivalent to the concept of “fitness landscape” [1,2]. The dynamic part of the selection originates from the competition among individuals. We assume in this schematization that the competition arises only

because the population shares some external resource that is evenly distributed, i.e., we disregard the effects of interindividual competition [3,4]. In this limit the effects of competition do not depend on the distribution of genotypes, and simply limit the total size of the population.

The most important effects of the selection can be roughly schematized by assuming that certain genotypes are forbidden, so that they are eliminated from the accessible space. Let us consider for the moment that evolution takes place in a flat fitness landscape. In this framework, there are no interactions among individuals: the evolution is given by the simple superposition of all possible lineages. The probability distribution of the population in the genotypic space at a given time can be obtained by summing over all possible individual “histories” in a way similar to the path integral formulation of statistical mechanics [5]. In this view, the phylogenetic lineage of an individual is given by the path connecting the genotypes of ancestors.

This assumption has to fail somewhere, since otherwise evolution would be equivalent to a diffusion process, without anything favoring the formation of species. It is often assumed that speciation events are rare and occur in a very short time [6], either due to some change in the fitness landscape (caused by an external catastrophe like the fall of a large meteorite, or by an internal rearrangement induced by coevolution) or because small isolated populations, escaping from competition, are free to explore their genotypic space (genetic drift) and find a path to a higher fitness peak [7]. Therefore, it is usually assumed that one can reconstruct the diverging time of two species from their genotypic distance from a single speciation event [8].

It is known, however, from recent investigations about the small-world phenomenon [9] that diffusion is seriously affected even by a small fraction of long-range jumps [10–12]. This fact could have dramatic consequences in our understanding of the large-scale evolution mechanism. If the short-range mutations are dominant, the evolution is equivalent to a diffusion process in the genotypic space. Assuming that the

\*Electronic address: bagnoli@dma.unifi.it

†Electronic address: bezzi@sisssa.it

fitness landscape is formed by mountains separated by valleys, and that the crests of the mountains are almost flat, then the large-scale evolution is dominated by the times needed to cross a valley by chance, while the short scale is dominated by the neutral exploration of crests [6,13].

Vice versa, if the long-range mutations are important (eventually amplified by the small-world mechanism), the time needed to connect two genotypes does not depend on their distance nor on the shape of the fitness landscape and the fitness maxima are quickly populated. Moreover, in this scenario, the speciation phenomenon should not be ascribed to the “discovery” of a preexisting niche, but rather to the formation of that niche in a given ecosystem due to internal interactions (coevolution) or external physical changes (catastrophes). After formation, the niche is quickly populated because of long-range mutations. In this framework, allopatric speciation [7] loses its fundamental importance, and sympatric speciation due to coevolution [14,15] becomes a plausible alternative.

In order to evaluate quantitatively the relevance of these speculations, we introduce an individual-based model of an evolving population. We assume that the genetic information (the genotype) of an individual is represented by a binary string  $\mathbf{x}=(x_1, x_2, \dots, x_L)$  of  $L$  symbols  $x_i=0,1$  (multilocus model with two alleles). In this way we are modeling haploid organisms, i.e., bacteria or viruses, or, more appropriately, more archaic, prebiotic entities. The choice of a binary code is not fundamental but certainly makes things easier. It can be justified by thinking of a purine-pyrimidine coding, or of “good” and “bad” alleles for genes (units of genetic information). In this second version, 0 represents a good gene and 1 a bad one. This is a sort of “minimal model” which is often used to model the evolution of genetic populations [14–18]. For bacteria, there are indications that major and minor codons work in such a way [19]. The genotypic space is thus a Boolean hypercube of  $L$  dimensions, and each individual sits on a corner of this cube, according to its genotype. The  $2^L$  corners of this hypercube represent all possible genotypes, which are at a maximum distance equal to  $L$ .

In the next section, we formalize the model and introduce the mathematical representation of mutation mechanisms. Then, in Sec. III, we consider the consequences of short- and long-range mutations for a flat fitness landscape. In the long-range case all genotypes are connected, regardless of their mutual distance: this case can be considered the equivalent of a mean-field approximation. We derive analytically the rate of spreading  $v$  and the characteristic spreading time  $\tau$  of a genetically homogeneous inoculum in the short-range ( $v_s, \tau_s$ ) and long-range ( $v_l, \tau_l$ ) cases. In the first case,  $v_s$  is independent of the genotype length  $L$ , and  $\tau_s$  grows linearly with  $L$ ; the opposite happens in the long-range case. Clearly, this different behavior has dramatic evolutionary consequences as  $L$  becomes large.

In general, however, only a small set of all possible long-range mutations is observed in real organisms. We thus compute numerically the rate of spreading  $v$  in a mixed short-range and sparse long-range case. A small-world effect can be observed: in the limit of large genotype lengths, a vanishing fraction of long-range mutations cooperates with the

short-range ones to give essentially the mean-field results.

In Sec. IV, we show that the effects of mutations and selection can be separated in the limit of a very smooth fitness landscape and mean-field long-range mutations. In this approximation we obtain analytically the result that the asymptotic probability distribution is a Boltzmann equilibrium one, in which the fitness plays the role of energy and mutations correspond to temperature. We compute numerically the asymptotic probability distribution for several fitness landscapes, finding that the equilibrium hypothesis is verified. As a nontrivial example of small-world effects in evolution, we checked numerically that this scenario still holds for a sparse long-range mutation matrix. Conclusions and perspectives are drawn in the last section.

## II. THE MODEL

In order to be specific, let us assume that the genotype of an individual is represented by a Boolean string of length  $L$ . In this way, the genotypic space is a Boolean hypercube with  $2^L$  nodes.

The genotype of an individual is represented by a string  $\mathbf{x}=(x_1, x_2, \dots, x_L)$  of  $L$  Boolean symbols  $x_i=0,1$ . Each position corresponds to a locus whose gene has two allelic forms. In this way we are modeling haploid (only one copy of each gene) organisms, i.e., bacteria or viruses, or, more appropriately, more archaic, prebiotic entities. The genotype can also be viewed as a spin configuration. In this case we use the symbol  $\sigma_i=2x_i-1$ .

We consider two kinds of mutation: point mutations, which interchange a 0 with a 1, and all other mutations, which do not alter the length of the genotype (transposition and inversions). We define the distance  $d(\mathbf{x}, \mathbf{y})$  between two genotypes  $\mathbf{x}$  and  $\mathbf{y}$  as the minimal number of point mutations needed to pass from  $\mathbf{x}$  to  $\mathbf{y}$  (Hamming distance):

$$d(\mathbf{x}, \mathbf{y}) = \sum_{i=1}^L (x_i - y_i)^2.$$

All possible genotypes of length  $L$  are distributed on the  $2^L$  vertices of a hypercube. A point mutation corresponds to a unit displacement on that hypercube (short-range jump).

The occurrence of point mutations in real organisms depends on the identity of the symbol and on its position on the genotype; in the present approximation, however, we assume that all point mutations are equally likely. Moreover, since the probability of observing a mutation is quite small, we impose the condition that at most one mutation is possible in one generation. The probability of observing a point mutation from genotype  $\mathbf{y}$  to genotype  $\mathbf{x}$  is given by the short-range mutation matrix  $M_s(\mathbf{x}, \mathbf{y})$ . Denoting by  $\mu_s$  the probability of a point mutation per generation, we have

$$M_s(\mathbf{x}, \mathbf{y}) = \begin{cases} 1 - \mu_s & \text{if } \mathbf{x} = \mathbf{y} \\ \frac{\mu_s}{L} & \text{if } d(\mathbf{x}, \mathbf{y}) = 1 \\ 0 & \text{otherwise.} \end{cases} \quad (1)$$

Other types of mutation correspond to long-range jumps in the genotypic space. The simplest approximation consists in assuming all mutations equiprobable. Let us denote by  $\mu_l$  the probability per generation of this kind of mutation. The long-range mutation matrix  $M_l(\mathbf{x}, \mathbf{y})$  is defined as

$$M_l(\mathbf{x}, \mathbf{y}) = \begin{cases} 1 - \mu_l & \text{if } \mathbf{x} = \mathbf{y} \\ \frac{\mu_l}{2^L - 1} & \text{otherwise.} \end{cases} \quad (2)$$

In the real world, only some kinds of mutation are actually observed. We model this fact by replacing  $M_l$  with a sparse matrix  $\hat{M}_l$ . We introduce the sparseness index  $s$ , which is the average number of nonzero off-diagonal elements of  $\hat{M}_l$ . The sum of these off-diagonal elements still gives  $\mu_l$ . In this case  $\hat{M}_l$  is a quenched sparse matrix, and  $M_l$  can be considered the average of the annealed version.

After considering both types of mutation, the overall mutation matrix is  $\mathbf{M} = \mathbf{M}_l \mathbf{M}_s$ , or  $\mathbf{M} = \hat{\mathbf{M}}_l \mathbf{M}_s$  for the quenched version.

We model our population at the level of the probability distribution of genotypes  $\mathbf{p} \equiv \mathbf{p}(t)$ , thus disregarding spatial effects. The evolution equation for  $\mathbf{p}$  is

$$p'(\mathbf{x}) = \frac{A(\mathbf{x}) \sum_{\mathbf{y}} M(\mathbf{x}, \mathbf{y}) p(\mathbf{y})}{\bar{A}}, \quad (3)$$

where the selection function  $A(\mathbf{x})$  corresponds to the average reproduction rate of individuals with genotype  $\mathbf{x}$ , and  $\bar{A} = \sum_{\mathbf{x}} A(\mathbf{x}) p(\mathbf{x})$  is the average reproduction rate of the population [1,2,20,21]. We write  $A(\mathbf{x})$  in an exponential form  $A(\mathbf{x}) = \exp[V(\mathbf{x})]$ , and we denote  $V(\mathbf{x})$  the *fitness landscape*.

The selection does not act directly on the genotype, but rather on the *phenotype* (how an individual appears to others). The phenotype of a given genotype can be interpreted as an array of morphological characteristics. We consider the simplest case, in which the phenotype is univocally determined by the genotype  $\mathbf{x}$ , which is not the general case, since polymorphism or age dependence is usually present. The general mapping between genotype and phenotype is largely unknown and is expected to be quite complex. The effects of some genes are additive (nonepistatic), while others can interact in a simple (control genes) or complex (morphologic genes) way.

A possible way of approximating these effects is to use the following form for the fitness  $V(\mathbf{x})$ :

$$V(\mathbf{x}) = \frac{\mathcal{H}}{L} \sum_{i=1}^L \sigma_i + \frac{\mathcal{J}}{L-1} \sum_{i=1}^{L-1} \sigma_i \sigma_{i+1} + \mathcal{K} \eta(\mathbf{x}), \quad (4)$$

where  $\eta(\mathbf{x})$  is a random function of  $\mathbf{x}$ , uniformly distributed between  $-1$  and  $1$  [ $\langle \eta(\mathbf{x}) \eta(\mathbf{y}) \rangle = \delta_{\mathbf{x}\mathbf{y}}$ ]. The ‘‘field’’ term  $\mathcal{H}$  represents a nonepistatic contribution to the fitness, in which all genes have equal weight. The ‘‘ferromagnetic’’ term  $\mathcal{J}$  represents simple interactions between pairs of genes (even though in general those are not symmetric) and corresponds

to a weakly rough landscape. Finally,  $\mathcal{K}$  modulates a widely rough landscape and can be thought of as an approximation of the effects of complex (spin-glass-like) interactions among genes.

### III. SPREADING AND SMALL-WORLD EFFECTS ON A FLAT FITNESS LANDSCAPE

In this section we study the case of a flat fitness landscape (no selection), i.e.,  $\mathcal{H} = \mathcal{J} = \mathcal{K} = 0$ . In this way we are modeling the evolution on the crest of a mountain, assuming that all deleterious mutations are immediately lethal (the part of the genotype that can originate this kind of mutation is not considered), and that we can neglect the small variation of fitness along the crest. This landscape is the one usually considered in the theory of neutral evolution [13]. Let us assume that this crest is colonized by a founder deme genotypically homogeneous (a  $\delta$  peak in the genotypic distribution). We want to obtain the average time needed to populate the crest according to the different mutation schemes. Since the fitness landscape is flat,  $\bar{A}$  is a constant.

We are interested in the behavior of  $\rho(t)$ , which is the average distance of the population from a given starting genotype  $\mathbf{x}_0$ , i.e.,

$$\rho(t) = \sum_{\mathbf{x}} p(\mathbf{x}, t) d(\mathbf{x}, \mathbf{x}_0). \quad (5)$$

We introduce the spreading velocity  $v$  as

$$v = \left. \frac{\partial \rho}{\partial t} \right|_{t=0}.$$

In the Appendix we obtain the spectral properties of the mutation matrices  $\mathbf{M}_s$  and  $\mathbf{M}_l$ . The corresponding spreading velocities in the limit of small time interval  $t$  compared to the characteristic spreading time  $\tau_s = L/2\mu_s$  (short range) and  $\tau_l = 1/\mu_l$  (long range) are  $v_s = \mu_s$ , Eq. (A3), and  $v_l = L\mu_l/2$ , Eq. (A4). If short- and long-range mutations coexist, one has  $v = v_s + v_l$ , Eq. (A5).

One can approximate the behavior of real biological systems by considering a mixture of short-range mutations, which occur with a relatively high frequency, and sparse long-range mutations, with sparseness index  $s$ .

We investigated the sparse case by numerical simulations, for some genotype lengths  $L$ . As shown in Fig. 1, as soon as the sparseness index  $s > 0$ , the numerical value of  $v$  becomes very close to the mean-field one, Eq. (A5). Notice that the average distance from the inoculum,  $\rho(t)$ , is rather insensitive to the distribution of  $\mathbf{p}$ . The actual distribution can be quite different from the one obtained with the mean-field matrix  $\mathbf{M}_l$ .

This transition may be interpreted as an indication of a small-world effect, and that there exists a first-order transition at  $s = 0$  [22]. However, the standard deviation of the spreading velocity,  $S(v)$ , appears to diverge at  $s = 0$  with an exponent  $1/2$ , as shown in Fig. 2.

These results suggest the presence of a small-world phenomenon in evolution: the rare and sparse long-range muta-

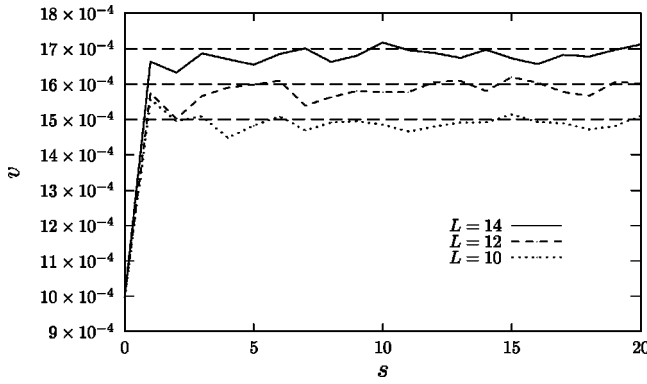


FIG. 1. The spreading velocity  $v$  vs sparseness  $s$  for three values of the genotype length  $L$ ;  $\mu_l=10^{-4}$ ,  $\mu_s=10^{-3}$ , flat fitness landscape. The dashed lines indicate the mean-field value  $v=\mu_s+L\mu_l/2$ . Average over 100 runs.

tions cooperate synergetically with the short-range ones to give essentially a mean-field effect. We checked that the previous results hold also for a nonflat (but smooth) fitness landscape ( $\mu_s=10^{-5}$ ,  $\mu_l=10^{-4}$  and  $\mathcal{H}=10^{-4}$ ,  $\mathcal{J}=0$ ,  $\mathcal{K}=0$ ). Again, as soon as  $s>0$ , the spreading velocity increases from the short-range values to the mean-field ones.

Our findings agree qualitatively with those in the literature [10,11], although we used a slightly different setup. Instead of rewiring links we added long-range jumps and our mutation matrix (which is the adjacency matrix of the problem) in general is not symmetric for the long-range part. Moreover, we are more interested in metric properties (the spreading velocity) than geometrical ones (the “chemical distance”).

#### IV. EQUILIBRIUM PROPERTIES AND SMALL-WORLD EFFECTS ON A SMOOTH FITNESS LANDSCAPE

In the following we shall study the effects of the cooperation between short-range and sparse long-range mutations on the equilibrium properties of the model.

Let us study the model in the presence of a smooth static

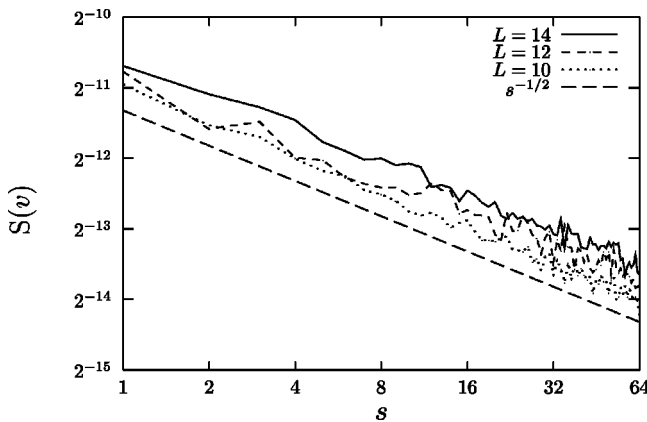


FIG. 2. The standard deviation  $S(v)$  of the spreading velocity  $v$  vs sparseness  $s$  for three values of the genotype length  $L$ ;  $\mu_l=10^{-4}$ ,  $\mu_s=10^{-3}$ , flat fitness landscape. The dashed line represents the law  $S(v)\sim s^{-1/2}$ . Average over 100 runs.

fitness landscape. In this case the fitness  $A$  does not depend explicitly on the distribution  $\mathbf{p}$ , and Eq. (3) can be linearized by using unnormalized variables  $z(\mathbf{x},t)$  satisfying

$$z(\mathbf{x},t+1)=\sum_{\mathbf{y}} A(\mathbf{y})M(\mathbf{x},\mathbf{y})z(\mathbf{y},t), \quad (6)$$

with the correspondence

$$p(\mathbf{x},t)=\frac{z(\mathbf{x},t)}{\sum_{\mathbf{y}} z(\mathbf{y},t)}.$$

In vectorial terms, Eq. (6) can be written as

$$\mathbf{z}(t+1)=\mathbf{M}\mathbf{A}\mathbf{z}(t), \quad (7)$$

where  $\mathbf{M}_{\mathbf{x}\mathbf{y}}=M(\mathbf{x},\mathbf{y})$  and  $\mathbf{A}_{\mathbf{x}\mathbf{y}}=A(\mathbf{x})\delta_{\mathbf{x}\mathbf{y}}$ .

When one takes into consideration only point mutations ( $\mathbf{M}\equiv\mathbf{M}_s$ ), Eq. (6) can be read as the transfer matrix of a two-dimensional Ising model [23–25], for which the genotypic element  $\sigma_i(t)$  corresponds to the spin in row  $t$  and column  $i$ , and  $z(\boldsymbol{\sigma},t)$  is the restricted partition function of row  $t$ . The effective Hamiltonian  $\mathcal{V}$  (up to constant terms) of a possible genealogical history  $\{\mathbf{x}(t)\}$  or  $\{\boldsymbol{\sigma}(t)\}$  from time  $1\leq t\leq T$  is

$$\mathcal{V}=\sum_{t=1}^{T-1}\left(\gamma\sum_{i=1}^L\sigma_i(t)\sigma_i(t+1)+V(\mathbf{x}(t))\right), \quad (8)$$

where  $\gamma=-\ln[\mu_s/(1-\mu_s)]$ .

This unusual two-dimensional Ising model has long-range coupling along the row (depending on the choice of the fitness function) and ferromagnetic coupling along the time direction (for small short-range mutation probability). In order to obtain the statistical properties of the system one has to sum over all possible configurations (stories), eventually selecting the right boundary conditions at time  $t=1$ . The bulk properties of Eq. (8) cannot be reduced in general to the equilibrium distribution of a one-dimensional system, since the transition probabilities among rows do not obey detailed balance. Moreover, the temperature-dependent Hamiltonian (8) does not allow an easy identification between energy and selection, and temperature and mutation, which is naively expected by the biological analogy with an adaptive walk.

#### A. Long-range mutations

Let us first consider the long-range mutation case. Equation (7), reformulated according to Eq. (3), corresponds to

$$\mathbf{z}(t+\tau)=(\mathbf{A}\mathbf{M}_l)^\tau\mathbf{z}(t).$$

Since it is easier to consider the effects of  $\mathbf{A}$  and  $\mathbf{M}_l$  separately, let us study in what limit they commute. The norm of the commutator on the asymptotic probability distribution  $\mathbf{p}$  is

$$\|[\mathbf{A}\mathbf{M}_l]\|=\sum_{\mathbf{xy}}|[\mathbf{A}\mathbf{M}_l]_{\mathbf{xy}}p(\mathbf{y})|,$$

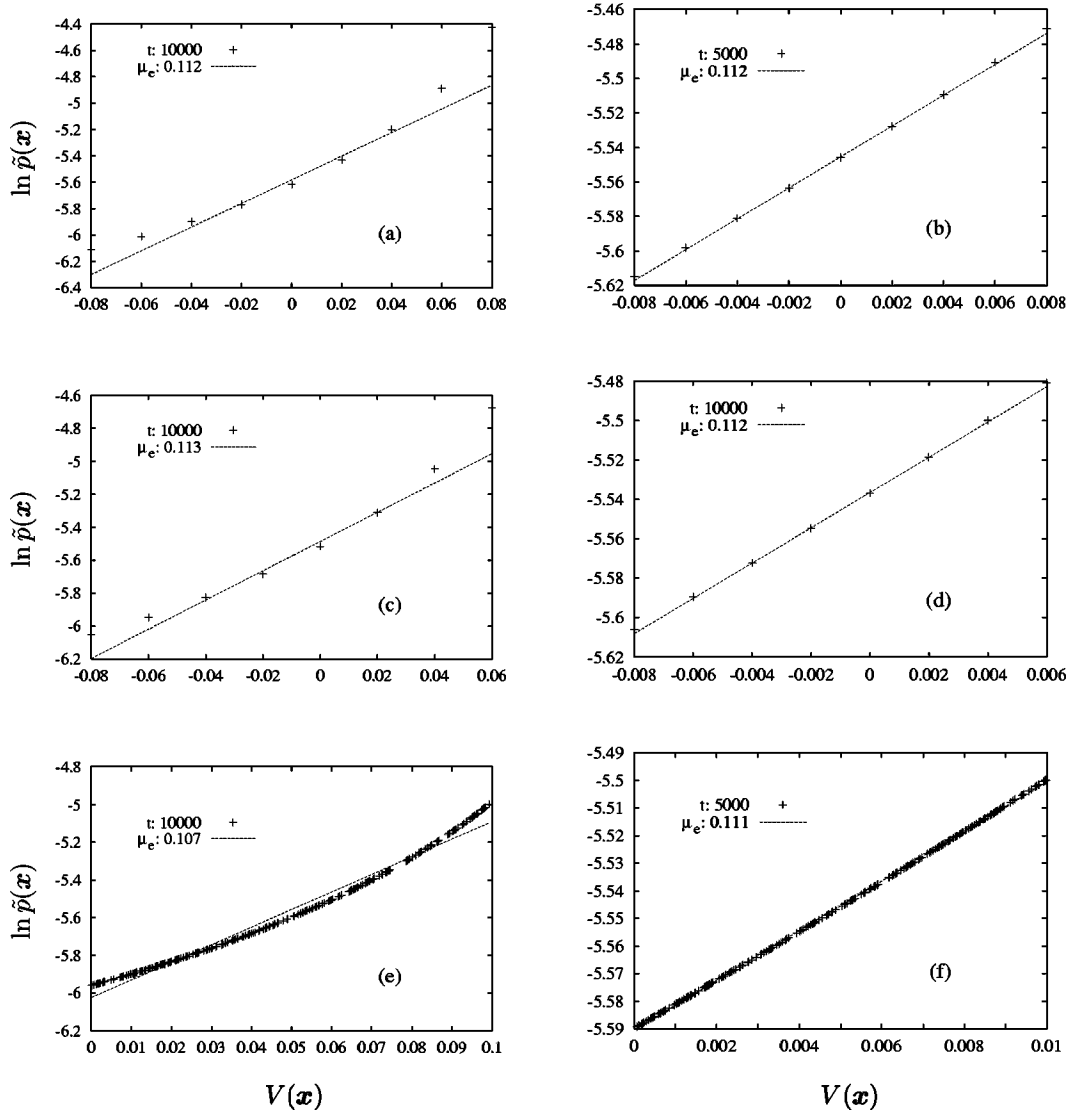


FIG. 3. Numerical check for long-range mutations. In the simulations we set  $L=8$ ,  $\mu_l=0.1$ , and  $\mu_s=0$ . We varied  $\mathcal{H}$  (a,b),  $\mathcal{J}$  (c,d), and  $\mathcal{K}$  (e,f), setting all other parameters to zero. (a)  $\mathcal{H}=0.01$ ,  $\mathcal{J}=0$ ,  $\mathcal{K}=0$ ; (b)  $\mathcal{H}=0.001$ ,  $\mathcal{J}=0$ ,  $\mathcal{K}=0$ ; (c)  $\mathcal{H}=0$ ,  $\mathcal{J}=0.01$ ,  $\mathcal{K}=0$ ; (d)  $\mathcal{H}=0$ ,  $\mathcal{J}=0.001$ ,  $\mathcal{K}=0$ ; (e)  $\mathcal{H}=0$ ,  $\mathcal{J}=0$ ,  $\mathcal{K}=0.1$ ; (f)  $\mathcal{H}=0$ ,  $\mathcal{J}=0$ ,  $\mathcal{K}=0.01$ . In the figures  $t$  indicates the number of generations,  $\mu_e$  the reciprocal of the slope of linear regression.

and it is bounded by  $\mu_l c$ , where  $c = \max_{xy} |A_{xx} - A_{yy}|$ . In the limit  $\mu_l c \rightarrow 0$  (i.e., a very smooth landscape), to first order in  $c$  we have

$$(\mathbf{A}\mathbf{M}_l)^\tau = \mathbf{A}^\tau \mathbf{M}_l^\tau + O(\tau^2 \mu_l c) \mathbf{A}^{\tau-1} \mathbf{M}_l^{\tau-1},$$

which is analogous to the Trotter product formula.

When  $\tau$  is of order  $1/\mu_l$ ,  $\mathbf{M}_l^\tau$  is a constant matrix with elements equal to  $1/2^L$ , and thus  $\mathbf{M}_l^\tau \mathbf{p}$  is a constant probability distribution. If  $\mu_l$  is large enough,  $(\mathbf{A}\mathbf{M}_l)^\tau = \mathbf{A}^\tau \mathbf{M}_l^\tau$ .

The asymptotic probability distribution  $\tilde{\mathbf{p}}$  is thus proportional to the diagonal of  $\mathbf{A}^{1/\mu_l}$ :

$$\tilde{p}(\mathbf{x}) = C \exp\left(\frac{V(\mathbf{x})}{\mu_l}\right), \quad (9)$$

i.e., a Boltzmann distribution with Hamiltonian  $V(\mathbf{x})$  and temperature  $\mu_l$ . This corresponds to the naive analogy between evolution and equilibrium statistical mechanics. In other words, the genotypic distribution is equally populated if the phenotype is the same, regardless of the genotypic distance since we used long-range mutations.

We have checked this hypothesis numerically, by iterating Eq. (3) for a time  $T$  large enough to be sure of having reached the asymptotic state. We plotted the logarithm of the probability distribution  $\tilde{p}(\mathbf{x})$  versus the value of the Hamiltonian  $V(\mathbf{x})$ . We computed the slope  $1/\mu_e$  of the linear regression. The quantity  $\mu_e$  is the effective “temperature” of the probability distribution according to the equilibrium hypothesis.

The results for the mean-field mutation matrix  $\mathbf{M}_l$  are shown in Fig. 3. We see that the equilibrium hypothesis is well verified in the limit  $\mu_l \gg c$ ; and that convergence is

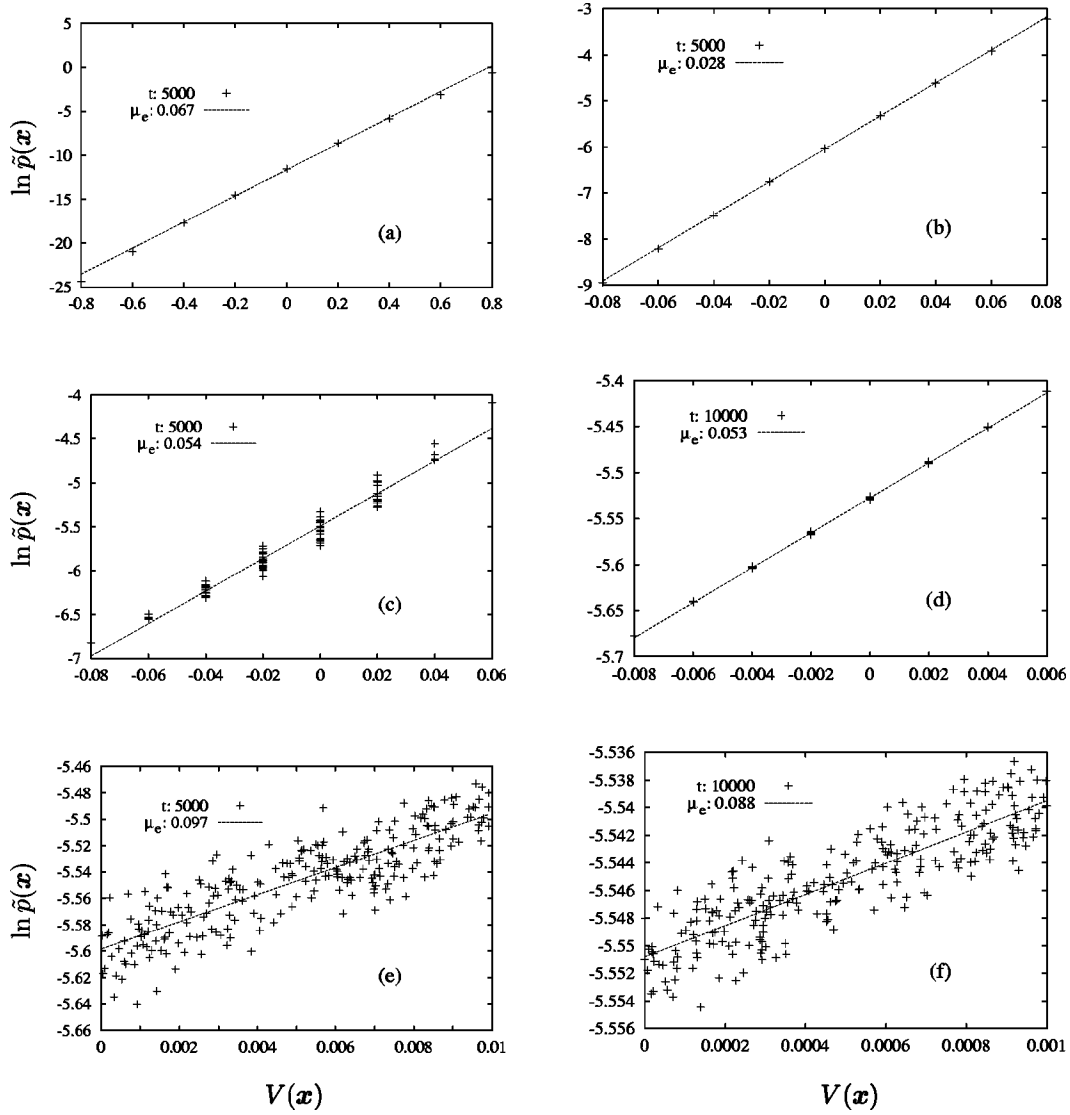


FIG. 4. Numerical check for short-range mutations. In the simulations we set  $L=8$ ,  $\mu_l=0$ , and  $\mu_s=0.1$ . We varied  $\mathcal{H}$  (a,b),  $\mathcal{J}$  (c,d), and  $\mathcal{K}$  (e,f), setting all other parameters to zero. (a)  $\mathcal{H}=0.1$ ,  $\mathcal{J}=0$ ,  $\mathcal{K}=0$ ; (b)  $\mathcal{H}=0.01$ ,  $\mathcal{J}=0$ ,  $\mathcal{K}=0$ ; (c)  $\mathcal{H}=0$ ,  $\mathcal{J}=0.01$ ,  $\mathcal{K}=0$ ; (d)  $\mathcal{H}=0$ ,  $\mathcal{J}=0.001$ ,  $\mathcal{K}=0$ ; (e)  $\mathcal{H}=0$ ,  $\mathcal{J}=0$ ,  $\mathcal{K}=0.01$ ; (f)  $\mathcal{H}=0$ ,  $\mathcal{J}=0$ ,  $\mathcal{K}=0.001$ . In the figures  $t$  indicates the number of generations,  $\mu_e$  the reciprocal of the slope of linear regression.

faster for a rough landscape. In all cases the effective temperature  $\mu_e$  is close to the expected value, i.e., to the mutation rate  $\mu_l$ .

### B. Short-range mutations

The above results hold only qualitatively for pure short-range mutations as shown in Fig. 4. The small dispersion of points in the figure implies that the genotypes can be divided into evenly populated groups sharing approximately the same fitness. This is always the case for the additive fitness landscape ( $\mathcal{H}$  contribution), since in this case the position of symbols in the genotype has no influence, and the short-range mutations are able to homogenize the distribution inside a group. This is only approximately valid for the  $\mathcal{J}$  contributions, since in this case the fitness depends on pairs of symbols, while mutations act only on single symbols.

However, as shown in Fig. 4(d), the homogeneous state is reached for a sufficiently high mutation probability. Finally, the homogeneous state is never reached for the very rough landscape case ( $\mathcal{H}$  contribution), even though the linear relation is satisfied on average.

In order to obtain a quantitatively correct prediction, one has to consider that the resulting slope  $\mu_e$  is related to the second largest eigenvalue  $\lambda_1$  of the mutation matrix by  $\mu = 1 - \lambda_1$ . When the fitness depends only on the “external field” term  $\mathcal{H}$ , in the asymptotic state the short-range mutations connect groups of equal fitness. This fact is reflected by the vanishing dispersion of points in Figs. 4(a) and 4(b).

In all cases, we expect that in the limit of a very smooth landscape  $\mu_e$  tends toward the limiting value  $2\mu_s/L$  of Eq. (A2). This limit is reached faster when only the additive  $\mathcal{H}$  term is present. This implies that for a large genotypes the effective temperature due to short-range mutations is

TABLE I. Effective temperature  $\mu_e$  for several values of the genotype length  $L$  and parameters of the fitness  $V(\mathbf{x})$ , Eq. (4). Here  $\mu_s=0.1$ ,  $\mu_l=0.01$ ; the row labeled  $\mathcal{H}$  stands for  $\mathcal{H}=0.1$ ,  $\mathcal{J}=\mathcal{K}=0$ , and the row labeled  $\mathcal{J}$  stands for  $\mathcal{J}=0.1$ ,  $\mathcal{H}=\mathcal{K}=0$ , the row labeled  $\mathcal{K}$  stands for  $\mathcal{K}=0.1$ ,  $\mathcal{H}=\mathcal{J}=0$ . We report the value of  $\mu_e^{(0)}$  for sparseness  $s=0$  (only small-range mutations) and the average value  $\bar{\mu}_e$  for  $s>s_c$ , where  $s_c$  is estimated visually from the plot of data (see also Fig. 5).

$V$	$L=6$			$L=8$			$L=10$			$L=12$		
	$\mu_e^{(0)}$	$\bar{\mu}_e$	$s_c$	$\mu_e^{(0)}$	$\bar{\mu}_e$	$s_c$	$\mu_e^{(0)}$	$\bar{\mu}_e$	$s_c$	$\mu_e^{(0)}$	$\bar{\mu}_e$	$s_c$
$\mathcal{H}$	0.035	0.045	3	0.025	0.035	3	0.020	0.030	2	0.017	0.027	2
$\mathcal{J}$	0.071	0.082	4	0.053	0.063	4	0.042	0.052	4	0.035	0.045	3
$\mathcal{K}$	0.080	0.105	6	0.090	0.105	6	0.096	0.107	6	0.098	0.111	3

vanishing.

In the opposite case, when the selection is strong, the application of the matrix  $\mathbf{A}$  “rotates” the distribution  $\mathbf{p}$  in a way that is essentially random with respect to the Fourier eigenvectors of  $\mathbf{M}_s$ . Thus, the effective second eigenvalue of the mutation matrix is given by  $1-\mu_s$ , obtained by averaging over all the eigenvalues of Eq. (A2). Consequently, we obtain  $\mu_e \approx \mu_s$ . This limit is reached faster in the case of a strongly disordered fitness landscape, i.e., in the limit of large  $\mathcal{K}$ . When only the  $\mathcal{J}$  term is present, one observes an intermediate case.

### C. Small-world effects

Let us now consider the case of a sparse long-range mutation matrix, coupled to a stronger short-range matrix, i.e.,  $\mu_s \gg \mu_l$ . We performed simulations with  $\mu_s=0.1$  and  $\mu_l=0.01$ ,  $L=6, 8$ , and  $10$ , and varying  $s$ . The results are summarized in Table I.

The effects of the two kinds of mutation are additive, implying that, for small  $\mu_l$ , the distributions are visually similar to the short-range case, Fig. 4. However, as soon as  $s$  is greater than a threshold  $s_c$ , the effective temperature  $\mu_e$  increases by the expected long-range contribution  $\mu_l$ . This increment is very relevant when the effect of short-range mutations is vanishing, i.e., for smooth landscapes ( $\mathcal{H}$  and  $\mathcal{J}$  contributions) and large genotypes. The effect is less relevant in the disordered case ( $\mathcal{K}$  contributions), since in this case the contribution to the effective temperature by the small-range mutations does not decrease with increasing  $L$ .

By increasing the weight of the sparse long-range mutations, the probability distribution also becomes similar to the long-range case, Fig. 3. We performed simulations with  $\mu_s=0.1$  and  $\mu_l=0.1$ ,  $L=6, 8, 10$ , and varying  $s$ . The results are shown in Table I. Notice that the value of  $s_c$  (estimated visually from the plot of data) is rather insensitive to  $L$  and the parameters of the fitness  $V(\mathbf{x})$ .

We quantified the distance between the resulting distribution and the Boltzmann one Eq. (9) through the computation of the average square difference from linear regression,  $\chi^2$ . In Fig. 5 we show that  $\chi^2$  becomes very small as soon as  $s > s_c$ , which seems to vanish with  $L \rightarrow \infty$ .

This implies that, for large genotypes, a vanishing fraction of long-range mutations, coupled to small-range ones, is sufficient to establish the statistical mechanics analogy of selection and mutations.

## V. CONCLUSIONS

We studied some simple models of asexual populations evolving on a smooth fitness landscape, in the presence of point mutations (small-range jumps in genotypic space) and other genetic rearrangements (long-range jumps).

We computed analytically the spreading velocity of an initially homogeneous inoculum on a flat fitness landscape, for the short-range and the long-range mean-field (all mutations equiprobable) cases. Since in a real situation only a small set of all possible mutations can occur, we also considered the quenched version of the long-range mutation matrix. In this case we showed that a small-world effect is present, since even a small number of quenched long-range jumps makes the results indistinguishable from those obtained by assuming all mutations equiprobable. These results still hold for a smooth fitness landscape.

We investigated this issue further, studying the equilibrium properties of the system in the presence of a smooth fitness landscape. In this framework, it was possible to show that the equilibrium distribution is a Boltzmann one, in which the fitness plays the role of an energy, and mutations that of a temperature. We checked this result numerically for different fitness landscapes, and a mean-field long-range mutation mechanism. As in the previous case, a small-world phenomenon appears, since similar results can be obtained

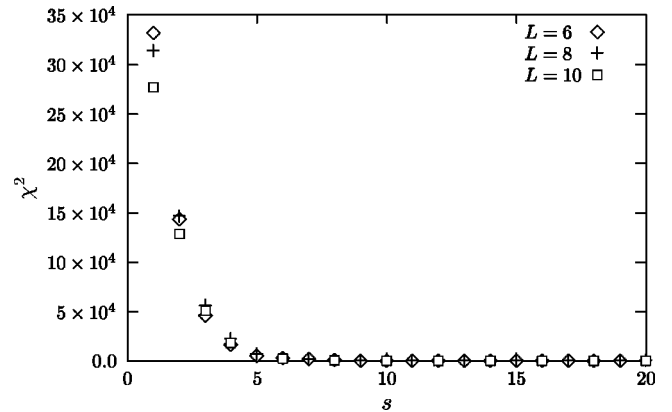


FIG. 5. Scaling of the average square difference from linear regression,  $\chi^2$ , vs sparseness  $s$ , for three values of the genotype length  $L$ ;  $\mu_l=0.1$ ,  $\mu_s=0.1$ ,  $\mathcal{H}=0$ ,  $\mathcal{J}=0$ ,  $\mathcal{K}=0.1$ . Averages taken over 20 runs.

using a combination of sparse long-range and short-range mutations.

We wish to acknowledge our participation in the Dynamics of Complex Systems [26] group. The numerical simulations were performed using the computational facilities of INFN PAIS 1999-G-IS-Firenze.

#### APPENDIX: SPECTRAL PROPERTIES OF THE MEAN-FIELD MUTATION MATRICES

Both  $\mathbf{M}_s$  and  $\mathbf{M}_l$ , Eqs. (1) and (2), are Markov matrices. Moreover, they are circular matrices, since the value of a given element does not depend on its absolute position but only on the distance from the diagonal. This means that their spectrum is real, and that the largest eigenvalue is  $\lambda_0=1$ . Since the matrices are irreducible, the corresponding eigenvector  $\xi_0$  is nondegenerate, and corresponds to the flat distribution  $\xi_0(\mathbf{x})=1/2^L$ . In analogy with circular matrices in the usual space, one can obtain their complete spectrum using the analog of Fourier transform in a Boolean hypercubic space. Let us define the ‘‘Boolean scalar product’’  $\odot$ :

$$\mathbf{x} \odot \mathbf{y} = \bigoplus_{i=1}^L x_i y_i,$$

where the symbol  $\oplus$  represent the sum modulus two (XOR) and the multiplication can be substituted by an AND (which has the same effect on Boolean quantities). This scalar product is obviously distributive with respect to the XOR:

$$(\mathbf{x} \oplus \mathbf{y}) \odot \mathbf{z} = (\mathbf{x} \odot \mathbf{z}) \oplus (\mathbf{y} \odot \mathbf{z}).$$

Note that the operation  $\mathbf{x} \oplus \mathbf{y}$  is performed bitwise between the two genotypes:  $(\mathbf{x} \oplus \mathbf{y})_i = x_i \oplus y_i$ .

Given a function  $f(\mathbf{x})$  of a Boolean quantity  $\mathbf{x} \in \{0,1\}^L$ , its ‘‘Boolean Fourier transform’’ is  $\tilde{f}(\mathbf{k})$  ( $\mathbf{k} \in \{0,1\}^L$ )

$$\tilde{f}(\mathbf{k}) = \frac{1}{2^L} \sum_{\mathbf{x}} (-1)^{\mathbf{x} \odot \mathbf{k}} f(\mathbf{x}).$$

The antitransformation operation is determined by the definition of the Kronecker delta

$$\delta_{\mathbf{k}\mathbf{0}} = \frac{1}{2^L} \sum_{\mathbf{x}} (-1)^{\mathbf{x} \odot \mathbf{k}},$$

and is given by

$$f(\mathbf{x}) = \sum_{\mathbf{k}} (-1)^{\mathbf{x} \odot \mathbf{k}} \tilde{f}(\mathbf{k}).$$

One can easily verify that the Fourier vectors  $\xi_k(\mathbf{x}) = (-1)^{\mathbf{x} \odot \mathbf{k}}$  are eigenvectors of both  $\mathbf{M}_l$  and  $\mathbf{M}_s$ , with eigenvalues

$$\lambda_0 = 1,$$

$$\lambda_{\mathbf{k}} = 1 - \mu_l - \frac{\mu_l}{2^L - 1} \quad (\text{A1})$$

for the long-range case, and

$$\begin{aligned} \lambda_0 &= 1, \\ \lambda_{\mathbf{k}} &= 1 - \frac{2\mu_s d(\mathbf{k}, \mathbf{0})}{L} \end{aligned} \quad (\text{A2})$$

for the short-range case, where  $d(\mathbf{x}, \mathbf{y})$  is the Hamming distance between genotypes  $\mathbf{x}$  and  $\mathbf{y}$ .

The computation of  $\rho(t)$ , Eq. (5), is easily performed in Fourier space, using the analog of the Parsifal theorem:

$$\begin{aligned} \sum_{\mathbf{x}} f(\mathbf{x}) g(\mathbf{x}) &= \sum_{\mathbf{x}} \sum_{\mathbf{k}} \sum_{\mathbf{k}'} \tilde{f}(\mathbf{k}) \tilde{g}(\mathbf{k}') (-1)^{\mathbf{x} \odot (\mathbf{k} \oplus \mathbf{k}')} \\ &= 2^L \sum_{\mathbf{k}} \sum_{\mathbf{k}'} \tilde{f}(\mathbf{k}) \tilde{g}(\mathbf{k}') \delta_{\mathbf{k}\mathbf{k}'} = 2^L \sum_{\mathbf{k}} \tilde{f}(\mathbf{k}) \tilde{g}(\mathbf{k}), \end{aligned}$$

where we have used the property  $\mathbf{k} \oplus \mathbf{k}' = \mathbf{0}$  if and only if  $k_i = k'_i$  for each component  $i$ . Let us denote by  $\mathbf{e}_n$  the unit vector along direction  $n$ , i.e.,  $(\mathbf{e}_n)_i = \delta_{ni}$ .

The Fourier transform of the distance  $d(\mathbf{x}, \mathbf{y})$  is obtained considering that  $(-1)^{(\mathbf{x} \oplus \mathbf{y}) \odot \mathbf{e}_n}$  gives 1 if  $x_n = y_n$  and  $-1$  otherwise; thus

$$d(\mathbf{x}, \mathbf{y}) = \frac{1}{2} \left( L - \sum_{n=0}^{L-1} (-1)^{(\mathbf{x} \oplus \mathbf{y}) \odot \mathbf{e}_n} \right).$$

We obtain

$$\begin{aligned} \tilde{d}_{\mathbf{x}_0}(\mathbf{k}) &= \frac{1}{2^L} \sum_{\mathbf{x}} d(\mathbf{x}, \mathbf{x}_0) (-1)^{\mathbf{x} \odot \mathbf{k}} \\ &= \frac{1}{2^L} \frac{1}{2} \sum_{\mathbf{x}} \left( L - \sum_{n=0}^{L-1} (-1)^{(\mathbf{x} \oplus \mathbf{x}_0) \odot \mathbf{e}_n} \right) (-1)^{\mathbf{x} \odot \mathbf{k}} \\ &= \frac{L}{2} \delta_{\mathbf{k}\mathbf{0}} - \frac{1}{2} \left( \sum_{n=0}^{L-1} (-1)^{\mathbf{x}_0 \odot \mathbf{e}_n} \right) \\ &\quad \times \left( \sum_{\mathbf{x}} (-1)^{\mathbf{x} \odot (\mathbf{k} \oplus \mathbf{e}_n)} \right) \\ &= \frac{L}{2} \delta_{\mathbf{k}\mathbf{0}} - \frac{1}{2} \sum_{n=0}^{L-1} (-1)^{\mathbf{x}_0 \odot \mathbf{e}_n} \delta_{\mathbf{k}\mathbf{e}_n}; \end{aligned}$$

i.e.,

$$\tilde{d}_{\mathbf{x}_0}(\mathbf{k}) = \begin{cases} L/2 & \text{if } \mathbf{k} = \mathbf{0} \\ -1/2 & \text{if } \mathbf{k} = \mathbf{e}_n \text{ and } (\mathbf{x}_0)_n = 0 \\ 1/2 & \text{if } \mathbf{k} = \mathbf{e}_n \text{ and } (\mathbf{x}_0)_n = 1 \\ 0 & \text{otherwise.} \end{cases}$$

The probability distribution  $p(\mathbf{x}, t)$  can be expanded on the eigenvector basis  $\xi_k(\mathbf{x})$  of  $\mathbf{M}$ :

$$p(\mathbf{x}, 0) = \sum_{\mathbf{k}} a_{\mathbf{k}} \xi_{\mathbf{k}}(\mathbf{x}) = \sum_{\mathbf{k}} a_{\mathbf{k}} (-1)^{\mathbf{x} \odot \mathbf{k}}$$



and

$$p(\mathbf{x}, t) = \mathbf{M}^t p(\mathbf{x}, 0) = \sum_{\mathbf{k}} a_{\mathbf{k}} \lambda_{\mathbf{k}}^t \xi_{\mathbf{k}}(\mathbf{x}),$$

i.e.,

$$\tilde{p}(\mathbf{k}, t) = a_{\mathbf{k}} \lambda_{\mathbf{k}}^t.$$

Thus

$$\begin{aligned} \rho(t) &= \sum_{\mathbf{k}} \tilde{d}_{\mathbf{x}_0}(\mathbf{k}) \tilde{p}(\mathbf{k}, t) = \sum_{\mathbf{k}} \tilde{d}_{\mathbf{x}_0}(\mathbf{k}) a_{\mathbf{k}} \lambda_{\mathbf{k}}^t \\ &= \frac{L}{2} \lambda_0^t + \sum_{n=0}^{L-1} \frac{1}{2} a_{\mathbf{e}_n} \lambda_{\mathbf{e}_n}^t. \end{aligned}$$

If at  $t=0$  the distribution is concentrated at  $\mathbf{x}_0 = \mathbf{0}$  [ $p(\mathbf{x}, 0) = \delta_{\mathbf{x}_0}$ ] then  $a_{\mathbf{k}} = 1$  for all  $\mathbf{k}$ . In both the short- and long-range cases,  $\lambda_{\mathbf{e}_n}$  does not depend on  $n$  [see Eqs. (A1) and (A2)], and thus

$$\rho(t) = \frac{L}{2} (\lambda_0^t - \lambda_1^t).$$

For the short-range case we have

$$\begin{aligned} \rho_s(t) &= \frac{L}{2} \left[ 1 - \left( 1 - \frac{2\mu_s}{L} \right)^t \right] = \frac{L}{2} \left\{ 1 - \exp \left[ t \ln \left( 1 - \frac{2\mu_s}{L} \right) \right] \right\} \\ &\simeq \frac{L}{2} \left[ 1 - \exp \left( -\frac{t}{\tau_s} \right) \right], \end{aligned}$$

and the characteristic spreading time is  $\tau_s = L/2\mu_s$ . For  $t$  small compared with  $\tau_s$  ( $L \rightarrow \infty$ ) we have

$$\rho_s(t) \simeq \mu_s t \equiv v_s t. \quad (\text{A3})$$

For long-range mutations we have

$$\rho_l(t) = \frac{L}{2} \left[ 1 - \left( 1 - \mu_l - \frac{\mu_l}{\mathbf{e}_n L - 1} \right)^t \right] \simeq \frac{L}{2} \left[ 1 - \exp \left( -\frac{t}{\tau_l} \right) \right]$$

with  $\tau_l = 1/\mu_l$ . For  $t \ll \tau_l$  ( $\mu_l \rightarrow 0$ ) we have

$$\rho_l(t) \simeq \frac{L\mu_l}{2} t \equiv v_l t. \quad (\text{A4})$$

The behavior of  $\rho(t)$  for short times, vanishing mutation probabilities, and large genotypes, Eqs. (A3) and (A4), is rather trivial. In these approximations one can neglect back mutations, and obtain the same results on a Cayley tree. However, the full analysis gives the exact behavior of  $\rho(t)$  for all times.

In the mixed case one has  $\mathbf{M} = \mathbf{M}_s \mathbf{M}_l$ . Since  $\mathbf{M}_s$  and  $\mathbf{M}_l$  share the same eigenvectors,  $\lambda = \lambda_s \lambda_l$  and, in the previous approximations,

$$v = v_s + v_l. \quad (\text{A5})$$

- 
- [1] R. A. Fisher, *The Genetical Theory of Natural Selection* (Dover, New York, 1930).
- [2] S. Wright, in *Proceedings of the 6th International Congress on Genetics, Ithaca*, edited by D.F. Jones (Brooklyn Botanical Garden, New York, 1936), pp. 356–366.
- [3] F. Bagnoli and M. Bezzi, *Phys. Rev. Lett.* **79**, 3302 (1997); e-print cond-mat/9708101.
- [4] F. Bagnoli and M. Bezzi, *Annu. Rev. Comput. Phys.* **8**, 265 (2000); e-print cond-mat/9906164.
- [5] F. Bagnoli and M. Bezzi, in *Path Integrals from peV to TeV*, edited by R. Casalbuoni *et al.* (World Scientific, Singapore, 1999), p. 493.
- [6] S. Gould and N. Eldredge, *Paleobiology* **3**, 114 (1997); *Nature* (London) **366**, 223 (1993).
- [7] E. Mayr, *Populations, Species and Evolution* (Harvard University Press, Cambridge, MA, 1970).
- [8] J. Maynard Smith, *Evolutionary Genetics* (Oxford University Press, Oxford, 1998).
- [9] D.J. Watts and S.H. Strogatz, *Nature* (London) **393**, 440 (1998).
- [10] M. Barthélemy and L. A. Nunes Amaral, *Phys. Rev. Lett.* **82**, 3180 (1999); **82**, 5180 (1999).
- [11] A. Barrat and M. Weigt, *Eur. Phys. J. B* **13**, 547 (2000).
- [12] R. Monasson, *Eur. Phys. J. B* **12**, 60 (1999).
- [13] M. Kimura, *The Neutral Theory of Molecular Evolution* (Cambridge University Press, Cambridge, 1983).
- [14] A. S. Kondrashov and F. A. Kondrashov, *Nature* (London) **400**, 351 (1999).
- [15] U. Dieckmann and M. Doebeli, *Nature* (London) **400**, 354 (1999).
- [16] W. Eigen, *Naturwissenschaften* **58**, 465 (1971); W. Eigen and P. Schuster, *ibid.* **64**, 541 (1977).
- [17] M. Doebeli, *J. Evol. Biol.* **9**, 893 (1999).
- [18] M. Doebeli, *J. Evol. Biol.* **9**, 893 (1999); U. Dieckmann and M. Doebeli, *Nature* (London) **400**, 354 (1999).
- [19] F. Bagnoli and P. Lió, *J. Theor. Biol.* **173**, 271 (1995).
- [20] D. L. Hartle, *A Primer of Population Genetics*, 2nd ed. (Sinauer, Sunderland, MA, 1988).
- [21] L. Peliti, in *Physics of Biomaterials: Fluctuations, Self-Assembly and Evolution*, edited by T. Riste and D. Sherrington (Kluwer, Dordrecht, 1996) pp. 287–308; e-print cond-mat/9505003.
- [22] M. Argollo de Menezes, C. F. Moukarzel, and T. J. P. Penna, *Europhys. Lett.* **50**, 5 (2000); e-print cond-mat/9903426.
- [23] I. Leuthäusser, *J. Stat. Phys.* **48**, 343 (1987).
- [24] P. Tarazona, *Phys. Rev. A* **45**, 6038 (1992).
- [25] T. Wiehe, E. Baake, and P. Schuster, *J. Theor. Biol.* **177**, 1 (1995).
- [26] <http://www.docs.unifi.it/>