J. theor. Biol. (1997) **184,** 51–64

Percolation on the Fitness Hypercube and the Evolution of Reproductive Isolation

SERGEY GAVRILETS†§ AND JANKO GRAVNER‡

† *Division of Environmental Studies, and the* ‡ *Department of Mathematics, University of California, Davis CA* 95616 *and the* §*Department of Mathematics, University of Tennessee, Knoxville, TN* 37996, *U.S.A.*

(*Received on* 1 *May* 1996, *Accepted in revised form* 20 *August* 1996)

We study the structure and properties of adaptive landscapes arising from the assumption that genotype fitness can only be 0 (inviable genotype) or 1 (viable genotype). An appropriate image of resulting (''holey'') fitness landscapes is a (multidimensional) flat surface with many holes. We have demonstrated that in the genotype space there are clusters of viable genotypes whose members can evolve from any member by single substitutions and that there are ''species'' defined according to the biological species concept. Assuming that the number of genes, n, is very large while the proportion of viable genotypes among all possible genotypes, p, is very small, we have deduced many qualitative and quantitative properties of holey adaptive landscapes which may be related to the patterns of speciation. Relationship between p and n determines two qualitatively different regimes: subcritical and supercritical. The subcritical regime takes place if p is extremely small. In this case, the largest clusters of viable genotypes in the genotype space have size of order n and there are many of such size; typical members of a cluster are connected by a single (''evolutionary'') path; the number of different (biological) species in the cluster has order n; the expected number of different species in the cluster within k viable substitutions from any its member is of order k. The supercritical regime takes place if p is small but not extremely small. In this case, there exists a cluster of viable genotypes (a ''giant'' component) that has size of order $2^n/n$; the giant component comes ''near'' every point of the genotype space; typical members of the giant component are connected by many evolutionary paths; the number of different (biological) species on the ''giant'' component has at least order n^2; the expected number of different species on the ''giant'' component within k viable substitution from any its member is at least of order kn. At the boundary of two regimes all properties of adaptive landscapes undergo dramatic changes, a physical analogy of which is a phase transition. We have considered the most probable (within the present framework) scenario of biological evolution on holey landscapes assuming that it starts on a genotype from the largest connected component and proceeds along it by mutation and genetic drift. In this scenario, there is no need to cross any ''adaptive valleys''; reproductive isolation between populations evolves as a side effect of accumulating different mutations. The rate of divergence is very fast: a few substitutions are sufficient to result in a new biological species. We argue that macroevolution and speciation on ''rugged'' fitness landscapes proceed according to the properties of the corresponding holey landscapes.

Introduction

What determines the number of species on earth? Why are there so many (or, perhaps, so few) of them? Could they all have evolved from a small number of (or even a single) ''protospecies''? What underlies inviability of hybrids between different species? How genetically different are different species? These are some of the most fundamental questions faced by evolutionary theory. We will try to get some insight into these and related questions about macroevolution combining the standard population genetics framework with methods developed in the percolation theory and the theory of random graphs.

First, one has to decide what a species is. There are many definitions and concepts of species. Here we shall use *the biological species concept* (Dobzhansky,

† Author to whom correspondence should be addressed. Present address: Department of Mathematics, University of Tennessee, Knoxville, TN 37996-1300, U.S.A. E-mail: gavrila@math.utk.edu

0022–5193/97/010051 + 14 $25.00/0/jt960242

1937; Mayr, 1942, 1963), which perhaps is the most common definition. According to this definition, species are groups of interbreeding natural populations that are reproductively isolated from other such groups. Evolution of reproductive isolation is influenced (at least potentially) by many genetical, ecological, developmental, behavioral, environmental, and other factors in different ways. If one wants to make the discussion less speculative, one should necessarily concentrate on only some of them while neglecting others. We will consider only *post-zygotic isolation* manifested in (and defined as) zero fitness of hybrids. Following most previous theoretical discussions of the evolution of post-zygotic isolation, we consider diploid populations under constant viability selection, assuming that the loci are diallelic, that the population is dioecious, that sexes are equivalent with respect to fitness, and that mating is random. Within this standard population genetics framework, an individual is represented by a combination of genes (i.e., its genotype) having some fitness.

Answers to the questions asked at the beginning of this paper depend on the *adaptive landscape* (Wright, 1931, 1980), i.e., the relation between genotype and fitness. Following Wright, adaptive landscapes are usually imagined as having many local "adaptive peaks" of different height separated by "adaptive valleys" of different depth. Adaptive peaks are interpreted as different species, adaptive valleys between them are interpreted as unfit hybrids (e.g., Barton, 1989); adaptive evolution is considered as local "hill climbing" (e.g., Kauffman & Levin, 1987). However, there are problems with this description and some of its implicit assumptions can be questioned. For instance, is it appropriate to assume that different species have different fitness? Small differences in fitness between individuals are important in *micro*evolution, but is this description appropriate for *macro*evolution? What is basically known is that there are some "good" combinations of genes representing fit individuals and "bad" combinations of genes representing unfit individuals (e.g., hybrids between different species). Microevolution, to a large extent, can be considered as an optimization problem, but is this so in the case of macroevolution? There are additional considerations coming from theoretical population genetics. Random genetic drift is increasingly important in multilocus systems (e.g., Gavrilets & Hastings, 1995). With random drift there can be practically no difference between survival probabilities of individuals with "deterministic" fitness 1 and fitness 0.9 or between survival probabilities of individuals with fitness 0.1 and fitness 0. Finally, there is a fundamental problem realized already by Wright. How can a population evolve from one local peak to another across an adaptive valley when selection opposes any changes away from the current adaptive peak? To solve this problem Wright (1931) proposed a (verbal) shifting-balance theory. Recent formal analyses of different versions of the shifting-balance theory (Lande, 1979, 1985; Barton & Rouhani, 1993; Rouhani & Barton, 1993; Gavrilets, 1996; Coyne *et al*., 1996) have shown that although the mechanisms underlying this theory can, in principle, work, the conditions are rather strict. Another possibility to escape a local adaptive peak is provided by founder effect speciation (Mayr, 1942, 1954; Carson, 1968; Templeton, 1980; Gavrilets & Hastings, 1996), but the generality of this scenario remains controversial. All these factors and considerations lead us to conclude that a different simplified description of adaptive landscapes may be both sufficient to get insight into the problem of speciation and even be more accurate as far as macroevolutionary phenomena are concerned.

The basic assumption made here is that fitnesses can take only two values: 1 (viable genotype) and 0 (inviable genotype). This description of adaptive landscapes is very closely related to the idea proposed by Dobzhansky almost 60 years ego (Dobzhansky, 1937). His original model considers a two-locus two-allele population initially monomorphic for a genotype, say **aaBB**. This population is broken up into two geographically isolated parts. In one part, mutation causes substitution of **a** for **A** and a local race **AABB** is formed. In the other part, mutation causes substitution of **B** for **b**, giving rise to a local race **aabb**. It is assumed that there is no reproductive isolation among genotypes **AABB, AaBB** and **aaBB** and among genotypes **aaBB**, **aaBb** and **aabb**, i.e., all offspring of matings within these two groups are viable. In contrast, genotypes **AABB** and **aabb** are considered to be reproductively isolated in the sense that double heterozygote **AaBb** is inviable. In this scheme, strong selection against hybrids between races with the genotypes **AABB** and **aabb** can be achieved, even though selection acting during the evolutionary divergence is weak or absent.

Dobzhansky's model implies that genotypes are of two types (viable and inviable) and that viable genotypes form "clusters" in genotype space so that the population can move from one viable state to another one separated by an adaptive valley following a "rim" or a "path" of viable genotypes without crossing any adaptive valleys. Populations diverge as a consequence of accumulation of different mutations (resulting from randomness of mutation and genetic drift) and reproductive isolation arises as a side effect

of these accumulating differences between populations. Founder events can increase the rate of divergence, but divergence will also happen in stable populations. Different properties of population genetic models utilizing the same idea have been discussed and formally studied (e.g., Nei *et al*., 1983; Bengtsson & Christiansen, 1983; Bengtsson, 1985; Barton & Bengtsson, 1986; Cabot *et al*., 1994; Wagner *et al*., 1994; Orr, 1995; Gavrilets & Hastings, 1996). In all these papers the existence of a chain of viable genotypes connecting two reproductively isolated genotypes was *postulated*. Below we will show that such chains (or clusters) of viable genotypes are expected under broad conditions.

We shall assign genotype fitnesses randomly. Random assignment of fitnesses often is used to get ideas about some "general" properties of population genetics models (e.g., Karlin & Carmelli, 1975; Lewontin *et al*., 1978; Ginzburg & Braumann, 1980; Turelli & Ginzburg, 1983). Properties of "rugged" landscapes with multiple peaks and valleys resulting from the assumption that fitnesses take any values between zero and one have been studied in a pioneering paper by Kauffman and Levin (1987) and in subsequent publications stimulated by that paper. The main purpose of our paper is similar to that one of Kauffman & Levin (1987). An appropriate three dimensional image of the fitnesses landscape we are interested in is a flat surface with a lot of holes like in a slice of Swiss cheese. Here we will study the structure of these "holey" landscapes resulting from the assumption that fitnesses take only values 0 and 1. A major difference of our approach, besides the assumption about possible fitness values and the techniques used, is that it focuses on the problem of speciation within the biological species concept. In contrast, the approach developed by Kauffman & Levin can be appropriate, in the strict sense, only if populations are asexual haploid.

Here each genotype will have a fixed probability, denoted by p, of being viable. Since p can also be considered as the probability of obtaining a viable genotype after combining genes randomly, it will be assumed very small (cf, Orr, 1995). The probability p can also be interpreted as a measure of environmental hostility: the smaller p is, the more difficult it is to survive. The probability p will be the same for all genotypes in some models and will vary among genotypes in other models. Under any form of random fitness assignment, viable genotypes generally will form sets in the genotype space connected by evolutionary paths. Connected sets of sites in multidimensional spaces are subject of *percolation theory* (e.g., Ballobás, 1985; Grimmett, 1989), whose terminology and methods we shall use. In the next section, we present several notions and definitions that will be used throughout the paper. After that we consider questions related to the maximum possible number of species in the whole genotype space. Then we discuss properties of "holey" landscapes arising when fitnesses are random. The last section summarizes our findings and discusses biological implications. An obvious limitation of our approach is the fact that we do not include any ecological factors.

Some Definitions

We consider diallelic loci whose number n typically will be very large. We shall use standard notation denoting alternative alleles at a locus with bold capital and lower-case letters and using w for fitnesses. A genotype formed by gametes i and j will be denoted as i/j. We shall consider two representations of the *genotype space*, i.e., the space of all possible genotypes. The first version is the most general. Each genotype is represented by a vertex of a $2n$-dimensional binary hypercube $B_n = \{0, 1\}^{2n}$. The location of a genotype on the i-th axes of B_n is determined by the number of alleles (0 or 1) represented by the corresponding capital letter at the i-th gene ($i = 1, 2, \ldots, 2n$). The overall number of genotypes in B_n is 4^n of which 2^n are homozygotes. An example of the genotype space B_n for a single locus case is given in Fig. 1(a). This representation allows for paternal-maternal and *cis-trans* effects, i.e., one-locus genotypes **A/a** and **a/A** are considered different, two-locus genotypes **AB/ab** and **Ab/aB** are considered different and so on. The second version of the genotype space implies that neither paternal-maternal nor *cis-trans* effects are present. Each genotype is represented by a "point" on a n-dimensional hypercube $Q_n = \{0, 1, 2\}^n$. The location of a genotype on the i-th axes is determined by the number of alleles (0, 1 or 2) at the i-th locus represented by the corresponding capital letter ($i = 1, 2, \ldots, n$). The overall number of genotypes in Q^n is 3^n of which 2^n are homozygotes. Examples of the genotype space B_n for one, two and three loci are given in Fig. 1(b–d). This representation of the genotype space is typical in population genetics models.

We will assume that *fitness* (viability) can take only two values: $w = 0$ (inviable genotype) and $w = 1$ (viable genotype). We will consider only *non-neutral* loci. An appropriate formal definition of a *neutral* locus is the following: locus A is neutral if

$$w(\mathbf{A}G/\mathbf{A}G') = w(\mathbf{A}G/\mathbf{a}G') =$$
$$w(\mathbf{a}G/\mathbf{A}G') = w(\mathbf{a}G/\mathbf{a}G')$$

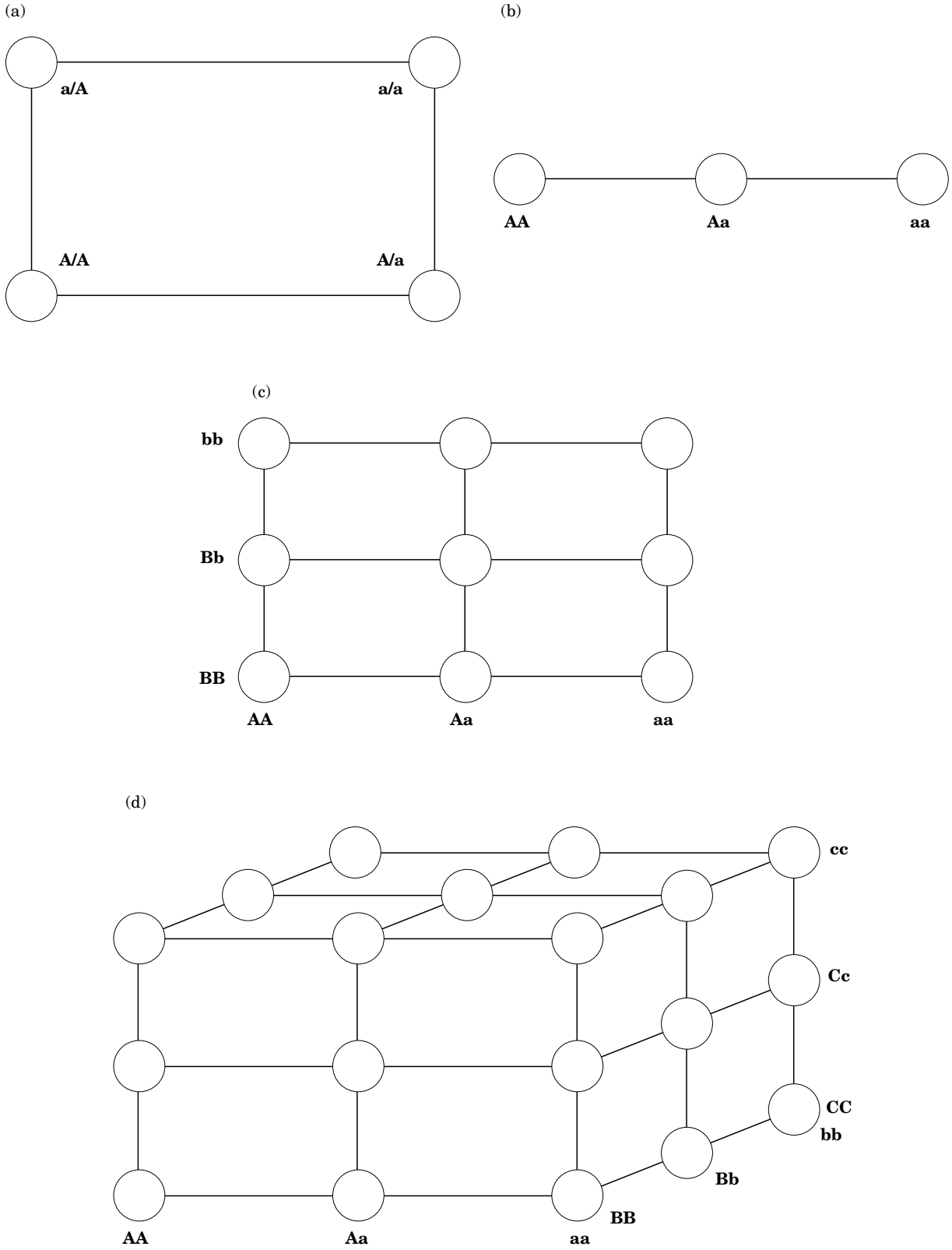

FIG. 1. Examples of genotype space. Genotype space B_n in the case of a single locus [part (*a*)]. Genotype space Q_n for $n = 1$, 2 and 3 [parts (*b*), (*c*) and (*d*), respectively]. In Fig. 1(d) only the genotypes on the "visible" side of the three-dimensional cube are shown.

for all genotypes G and G' in the remaining loci. This definition reflects the idea that no changes in a neutral locus affect fitness.

An *offspring* of a mating between two viable genotypes is any genotype that can be produced as a result of segregation and recombination. Two groups of viable genotypes will be considered as representing different *species* if all offspring resulting from matings "within" a group are viable and all offspring resulting from matings "between" groups are inviable. For example, in Dobzhansky's model two different species are represented by genotypes **AABB** and **aabb**.

A group of viable genotypes forming a species can, in principle, include any number of genotypes. We will be mainly interested in questions related to the number of "biological" species. For this purpose, it is sufficient to concentrate on only "monomorphic" species represented by a single homozygous genotype. This follows from a simple fact that although a "polymorphic" species includes several homozygotes and heterozygotes, there is no reproductive isolation among them. Thus, a polymorphic species contributes one and only one "monomorphic" species to the count of "monomorphic" species. A pleasant consequence of this property is that all complications introduced by recombination are avoided without any loss of generality.

A sequence of viable genotypes $x_0, x_1, \ldots, x_N$ is an *evolutionary path* if genotypes x_{i-1} and x_i are different in only a single gene. This means that genotype x_0 can evolve through viable genotypes into genotype x_N through fixation of consecutive mutations at a single locus. For example, in Dobzhansky's model the evolutionary path connecting genotypes **AAbb** and **aaBB** is **AAbb**, **AABb**, **AABB**, **AaBB**, **aaBB**. For any viable genotype x, the *connected component* of x is the set of all genotypes connected to x by an evolutionary path.

We will denote by L_1 the connected component with the largest number of homozygotes and by L_2 the component with the second largest number of homozygotes. We shall denote the *size* of L_i, i.e., the number of homozygotes in L_i, as $|L_i|$, and the number of species in L_i as N_i. Note that the number of species in a connected component is bounded from above by the number of homozygotes in this component, i.e., $N_i \leqslant |L_i|$. We will consider the most probable (within the present framework) scenario of biological evolution assuming that it starts on a genotype from the largest connected component and proceeds along it by mutation and genetic drift.

The graph-theoretical distance between two genotypes that belong to the same connected component is the length of the shortest evolutionary path connecting them. The *Hamming distance* between two genotypes is the number of genes in which these genotypes differ. For example, the Hamming difference between two two-locus genotypes **AABB** and **aabb** is four. In Dobzhansky's model, this is also the graph-theoretical distance.

Throughout the paper, a statement as "an event happens asymptotically" means that the probability that the event happens converges to 1 as the number of loci, n, becomes larger and larger.

Maximum Number of Species

We start by assuming that fitnesses can be assigned in an arbitrary way. Two interesting questions arise in this context. The first is about the maximum possible number of different species interconnected by evolutionary paths. The second is about the minimum possible proportion of viable genotype that makes these species connected by evolutionary paths.

We will consider maximum number of homozygous species different in at least n_{min} loci, denoting their number as $N(n, n_{min})$ and the minimum proportion of viable genotypes that connect them as $\pi(n, n_{min})$. For example, if $n_{min} = 2$ and there are only two loci, then the maximum number of species in Q_n is two and the minimum proportion of viable genotypes is 5/9, while if there are three loci, $N = 4$ and $\pi = 11/27$ (see Fig. 2). Several more general cases can be treated analytically (see the Appendix). For instance,

$$N(n, 2) = 2^{n-1},\ \pi(n, 2) = (3\cdot 2^{n-1} - 1)/3^n, \quad (1a)$$

$$N(n, 3) = 2^{n-2},\ \pi(n, 3) = (3\cdot 2^{n-2} - 1)/3^n. \quad (1b)$$

For example, if $n_{min} = 2$ and $n = 100$, $N \approx 6\cdot 10^{29}$, $\pi \approx 4\cdot 10^{-18}$. If n_{min} is fixed, while n becomes very large, then asymptotically

$$C_1\cdot n^{-(n_{min}-2)}\cdot 2^n \leqslant N(n, n_{min}) \leqslant C_2\cdot n^{-(n_{min}-1)/2}\cdot 2^n, \quad (1c)$$

where C_1 and C_2 depend on n_{min}, but not on n. The fact that any pair of genotypes can be connected by an evolutionary path of at most $2n + 1$ viable genotypes immediately implies that

$$N(n, n_{min}) \leqslant 3^n\cdot\pi(n, n_{min}) \leqslant 2n\cdot N(n, n_{min}), \quad (1d)$$

where $3^n\cdot\pi(n, n_{min})$ is the number of genotypes forming evolutionary paths. Let n_{min} be a positive proportion of n, say $n_{min} = \alpha n$ for some $0 < \alpha < 1$. If $\alpha > 1/2$, then $N(n, n_{min}) \leqslant 2\alpha/(2\alpha - 1)$, so that N does not grow at all, and π decreases as $n\cdot 3^{-n}$ with increasing n. On the other hand, if $\alpha < 1/2$, then N increases exponentially. For example, it is known that the number of $0.1n$-separated genotypes with n loci is for large n between $e^{0.386n}$ and $e^{0.481n}$, and so is [by virtue

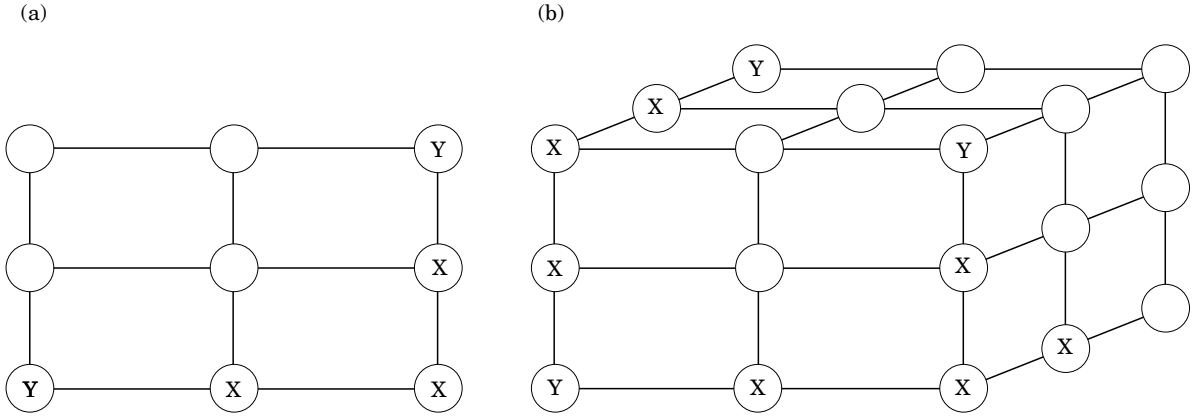

FIG. 2. Maximum number of homozygous species different in $n_{min} = 2$ loci. Biological species are marked by **Y**; viable genotypes forming evolutionary paths are marked by **X**; all other genotypes are inviable. (a) Genotype spaces Q_2: two biological species; overall number of viable genotypes is five (out of 9). (b) Genotype spaces Q_3: four biological species; overall number of viable genotypes is 11 (out of 27). Only the genotypes on the "visible" side of the three-dimensional cube are shown.

of (1d)] the number of genotypes necessary to connect them. These, and many other, asymptotic bounds can be found in Chapter 17 of MacWilliams & Sloane (1977) and Chapter 9 of Conway & Sloane (1988).

Common sense suggests that the proportion of viable genotypes among all possible genotypes should be very small while the number of evolutionary connected species may be large. To have a large number of species, one should have a lot of viable homozygotes and just enough viable heterozygotes to form large connected sets. It is perhaps not too surprising that one can construct a deterministic model (such as the one above) in which this happens. In the following sections, we show how the same phenomenon may be exhibited if adaptive landscapes are constructed *randomly*.

Random Fitnesses

Results presented below will have different degrees of generality and mathematical strictness. As often happens, the most complete analytical results are derived for the least plausible model which we consider in the next section.

EQUAL PROBABILITIES TO BE VIABLE: PERCOLATION ON B_n

In this model, each genotype in B_n is viable with probability $p > 0$ and is inviable with probability $1 - p$ independently of other genotypes. Consideration of the genotype space B_n implies that any change in the genotype including the flip of genes in a heterozygous locus (that is a change from **Aa** to **aA**) results in a completely independent fitness value. With large number of loci the overall number of viable genotypes is approximately $p \cdot 4^n$. Among those there are approximately $p \cdot 2^n$ homozygotes and $p \cdot 2^n(2^n - 1)$ heterozygotes. The heterozygote/homozygote ratio is approximately 2^n. Any large connected component of B_n has approximately the same heterozygote/homozygote ratio. If $p > 1/2$, then all viable genotypes are connected with probability approaching one, i.e., there is a single component (Burtin, 1977; Erdös & Spencer, 1979). Biologically that would mean that all genotypes could evolve from any single genotype without crossing any adaptive valleys. It can be shown (see Bollobás & Thomason, 1985), that for p values close to one, the number of species in this component is of order $-n/\log_2(1-p)$, hence of order $O(1)$ when $p \approx 1 - 2^{-n}$.

However, as was discussed above, it is more realistic to assume p to be small. We will scale the probability that a genotype is viable with the number of loci n,

$$p = \lambda/n, \tag{2}$$

where λ is allowed to depend on n.

Result 1(*a*): *number of homozygotes in largest connected components*

Asymptotically, if $\lambda > 1/2$, then for some positive functions α and β of λ

$$|L_1| > \alpha \cdot n^{-1} \cdot 2^n, |L_2| \leqslant \beta \cdot n, \tag{3a}$$

while if $\lambda < 1/2$,

$$|L_1| \leqslant \beta \cdot n. \tag{3b}$$

In the first case, when the proportion of viable genotypes is bigger than the critical value $1/(2n)$, there exist a "giant" component that includes a positive proportion of all viable homozygotes (the number of which is order $n^{-1} \cdot 2^n$ because p is order $1/n$). The second largest component has a much smaller size, order n. In percolation theory, this is usually referred

to as the *supercritical regime*. In the second case when the proportion of viable genotypes is smaller than the critical value, no connected component has size bigger than order n. This case is referred to as *subcritical regime*. Thus, we have just shown that large connected components whose existence was *postulated* in Dobzhansky-type models are expected to exist in this model even if the overall probability to get a viable genotype is very small [e.g., of order $O(1/n)$]. At the boundary between the sub-critical and supercritical regimes the system undergoes a *phase transition*, in the sense that the number of homozygotes in the largest component experiences a jump from order n to order $2^n/n$.

To understand percolation in high dimensions, such as percolation on B_n, one thinks of B_n as approximately a regular tree T in which each node (genotype) has $2n$ neighbors, the same as the number of neighbors in the genotype space. Assume that every node in T is independently "viable" with probability p, and "inviable" with probability $1-p$. Since T is now an infinite graph, we say that percolation occurs if there is a infinite path in the graph T, consisting entirely of "viable" nodes. The point is that it is very easy to see exactly when percolation occurs: just choose a viable node (a root) $x_0 \in T$ and observe that the probability of an infinite path started at x_0 is the same as the probability of survival of the branching process with expected number of successors equal to $p(2n-1)$. This shows that percolation occurs if $p > 1/(2n-1)$ and all paths are finite if $p \leqslant 1/(2n-1)$. The fact that the critical probability for existence of long paths is roughly $1/(2n)$ remains true for B_n. The tree comparison is useful for many other questions as well.

The next result describes what happens with the number of species. It shows that even when the overall proportion of viable genotypes in the genotype space is very small, say of order $O(1/n)$, the number of species in a connected component can be as large as $O(n^2)$.

Result 1(*b*): *number of species in the largest connected component*

Asymptotically if $\lambda > 1/2$,

$$\alpha_2 n^2/\lambda < N_1 < 2n^2/\lambda,\ N_2 < \beta_2 n, \tag{4a}$$

while if $\lambda < 1/2$,

$$\gamma_2 n/\ln(1/(2\lambda)) < N_1 < n/\ln(1/(2\lambda)). \tag{4b}$$

Here α_2, β_2, γ_2 are some positive functions of λ.

To interpret this result, let us think of n as fixed, and λ as starting at n (so that $p \approx 1$) and continuously decreasing towards zero. The number of species in the giant component then starts at one, quickly increases to order n when $\lambda \approx (1-\epsilon)n$ with some $\epsilon > 0$, and then increases steadily until λ is order 1 (and p is order $1/n$), when N_1 becomes of order n^2. Then, when λ is about $1/2$, i.e., at the point of phase transition, the number of species in the largest connected component suddenly drops down to order n. After that it continues to decrease slowly (being for example of order $n/\ln(n)$ when $\lambda \approx 1/n$). Figure 3 illustrates these features. Note that besides the point of phase transition at $p \approx 1/(2n)$, the number of species undergoes a dramatic change in the neighborhood of $p = 1$.

Result 1(*c*): *the geometric structure of* L_1

The supercritical regime. If $\lambda > 1/2$,

(i) the expected proportion of points in B_n within the Hamming distance 2 from L_1 converges to 1 as $n \to \infty$, while the probability that every point in B_n is within Hamming distance 3 from L_1 converges to one. This means that in the supercritical regime, L_1 comes "near" every point of B_n;

(ii) typical members of L_1 are connected by a large number of evolutionary paths: for every two genotypes x and y, and any positive integer k, asymptotically there are at least k disjoint evolutionary paths connecting x and y;

(iii) the Hamming distance and the graph-theoretical distance between two typical points on L_1 have the same order. This means that typical points on L_1 can be connected by evolutionary paths that are not extremely windy. At the same time, points that are close to each other (in the Hamming distance) may be connected by evolutionary paths that are much longer than the Hamming distance.

(iv) for large k (for example, for $k = \alpha n$) the number of different species in L_1 within k viable substitutions from x_0 is of the order $k \cdot n$, i.e. is very large.

The subcritical regime. If $\lambda < 1/2$,

(i) there is a large number of connected components of size $O(n)$. A typical point of B_n will be at the Hamming distance $O(n)$ from the closest one of these components. The connected components are very "thin";

(ii) typical members of L_1 are connected by a single evolutionary path;

(iii) asymptotically, the ratio of the Hamming distance and the graph-theoretical distance

between two typical points on a connected component is one.

(iv) for large k (for example, for $k = \alpha n$) the number of different species in L_1 within k viable mutations from x_0 is of the order k, i.e. is very small.

Let us consider a genotype x_0 on the giant component. Eventually, all genotypes in L_1 can evolve from x_0, but how fast? Let I_j be the probability that the first speciation event happens after substitution number j. Dependence of this probability on the number of loci n is suppressed in the notation; the inequalities below are valid in the limit as $n \to \infty$.

Result 1(*d*): *the rate of speciation on* L_1

Both in the supercritical and subcritical regimes, the first speciation event happens after substitution number j with probability I_j which is bounded by

$$1 - 2^j/\binom{2j}{j} \leqslant I_j \leqslant 1 - (e^{-\lambda}\lambda)^{2j} \cdot 2^j/\binom{2j}{j}.$$

For example, $I_2 \geqslant 1/3$ and $I_3 \geqslant 3/5$. In general, I_j is bounded below by a number which goes, very fast, to one as j increases (see Fig. 4). This result shows that speciation is an inevitable consequence of accumulating different mutations (cf., Orr, 1995).

EQUAL PROBABILITIES TO BE VIABLE: PERCOLATION ON Q_n

In this model, each genotype in Q_n is viable with probability $p > 0$ and is inviable with probability $1 - p$ independently of other genotypes. The model implies that flips of genes in heterozygous loci (that is change from **Aa** to **aA** etc.) do not change fitness value or, in biological terms, that paternal-maternal and cis-trans effects are absent. With large number of loci the overall number of viable genotypes is approximately $p \cdot 3^n$. Among those there are approximately $p \cdot 2^n$ homozygotes and $p \cdot (3^n - 2^n)$ heterozygotes. The heterozygote/homozygote ratio is approximately $(3/2)^{-n}$. Any large connected component of Q_n has approximately the same heterozygote/homozygote ratio. It can be shown that if $p > 2 - \sqrt{2}$, then all viable genotypes are connected with probability approaching one.

As before we will assume that p is very small and use the scaling (2).

Result 2(*a*): *number of homozygotes in largest connected components*

Asymptotically if $\lambda > 1$,

$$|L_1| > \alpha \cdot n^{-1} \cdot 2^n, |L_2| \leqslant \beta \cdot n, \tag{5a}$$

if $\lambda < 1/2$,

$$|L_1| \leqslant \beta \cdot n, \tag{5b}$$

and if $\lambda \in (1/2, 1)$,

$$2^{\alpha_1 n} < |L_1| < 2^{\alpha_2 n}. \tag{5c}$$

Result 2(*b*): *number of species in the giant component*

If $\lambda > 1$, then asymptotically

$$\alpha \cdot \sqrt{\frac{n}{p}} \leqslant N_1 \leqslant \frac{1}{2} \cdot \left(\frac{n}{p}\right)^2, \tag{6}$$

where α is a positive constant.

The boundaries on the number of species that we have been able to find are much broader for this model than in the previous one. However, some additional considerations make the following conjecture very plausible.

Conjecture B. In the supercritical regime the number of species N_1 is order n/p.

Results 2(*c*) *and* (*d*)

The geometrical structure of L_1 in Q_n and the rate of speciation on L_1 are similar to those of L_1 in B_n described in Result 1(c) and (d).

Results 2(a)–(d) show that qualitative properties of holey landscapes in this model (such as existence of connected components, existence of two drastically different regimes, phase transition at the boundary of these regimes etc.) are similar to those in the model considered in the previous section.

MULTIPLICATIVE MODEL: PERCOLATION ON Q_n

In this model, each genotype in Q_n is viable with a probability that decreases geometrically with number of heterozygous loci, i.e., each genotype is assigned fitness 1 with probability

$$p \cdot a^{\#\text{ of heterozygous loci}}.$$

Here a is a constant between zero and one and p is a small value which can depend on n. This model is an analog of the multiplicative model in population genetics. As before the model implies that paternal-maternal and cis-trans effects are absent. Note that the expected number of viable genotypes in this model is $p(2 + a)^n$ and, thus, the proportion of viable genotypes in the genotype space is approximately $p(2 + a)^n/3^n$ and extremely small.

We will scale the probability that a genotype is viable with the square root of the number of loci, n,

$$p = \lambda/\sqrt{n} \tag{7}$$

where λ is allowed to depend on n. Since homozygotes

have a large fitness advantage, one can expect that most of the evolutionary action happens in relatively small neighborhoods of homozygotes; next theorem makes this more precise.

Result 3(*a*): *number of homozygotes in the largest component*

Asymptotically if $\lambda > 1/\sqrt{a}$, then

$$|L_1| > \alpha \cdot n^{-1} \cdot 2^n \tag{8a}$$

and if $\lambda < (1/2)a \cdot \ln(1/a)$, then

$$|L_1| < \beta \cdot n \cdot \ln(n). \tag{8b}$$

Here α and β are some positive functions. Note that for the giant component to exist, p (the probability of survival for homozygotes) should be much bigger (order $1/\sqrt{n}$) than in the previous models (where it was order $1/n$).

Next theorem establishes that the number of species in the supercritical regime is exponentially large in this case.

Result 3(*b*): *number of species in the giant component*

The number of species in the supercritical regime is exponentially large in the sense that there exist a constant $\alpha_1 > 0$ such that

$$N_1 > e^{\alpha_1 n}. \tag{9}$$

This shows that the multiplicative model can provide an enormously large number of species while keeping the proportion of viable genotypes extremely small (cf. the section "Maximum number of species"). Although we are not able to give more precise exponential asymptotics for the number of species, we can conclude that in this case the largest number of species in a component drops even more dramatically between the supercritical and sub-critical regime: from exponential in n to the size at most $n \cdot \ln(n)$ (which is the number of homozygotes in the largest component). While presumably many geometrical aspects of L_1 remain the same as in the previous two models, this aspect of the multiplicative model remains largely unclear. However, we still expect that qualitative properties of holey landscapes in the multiplicative model are similar to those in the models considered in the previous sections.

Discussion

The standard methodology of theoretical population genetics is to analyze the dynamics of gene frequencies assuming some relationship between fitness and genotype (i.e., assuming some adaptive landscape). This approach does not allow to study questions related to macroevolution which depend on the structure of adaptive landscape over the whole genotype space. Our paper represents an attempt to analyze the structure and properties of typical fitness landscapes in some general models.

A widely accepted picture of adaptive landscapes, which goes back to Wright (1931), is the one with many adaptive peaks of different height separated by adaptive valleys of different depth. However, this representation of adaptive landscape has limitations discussed at the beginning of this paper (see also Whitlock *et al*., 1995). Here we have studied a different family of adaptive landscapes, which can be traced to a model proposed by Dobzhansky (1937). The basic assumption underlying our approach is that genotypes can be only of two types: viable and inviable. An appropriate image of resulting fitness landscapes is a flat surface with many holes. We feel that these "holey" adaptive landscapes may be a more appropriate model for studying patterns of speciation and macroevolution. Note that assuming fitnesses to be 0's and 1's does not contradict observations of intermediate values since these observations are averages over genetic background. Thus, this assumption is applicable to much more general settings than it might initially appear (Lev Ginzburg, personal communication).

Starting with the only assumption that there are "good" and "bad" combinations of genes we have demonstrated that in the genotype space (i) there are clusters (connected components) of viable genotypes whose members can evolve from any member by single mutations and drift, and (ii) there are species defined according to the biological species concept.

Previously existence of clusters of viable genotypes with different biological species was postulated (Dobzhansky, 1937; Nei *et al*., 1983; Bengtsson & Christiansen, 1983; Bengtsson, 1985; Barton & Bengtsson, 1986; Cabot *et al*., 1994; Wagner *et al*., 1994; Orr, 1995; Gavrilets & Hastings, 1996). In contrast, we have shown it to be expected under broad conditions.

Making two additional assumptions that the number of genes is very large while the proportion of "good" combinations of genes is very small, we have deduced many qualitative and quantitative properties of adaptive landscapes which may be related to the patterns of speciation. Depending on the relationship between the proportion of viable genotypes among all possible genotypes, p, and the number of loci, n, there can be two qualitatively different regimes: subcritical and supercritical.

The subcritical regime takes place if the proportion of viable genotypes is extremely small. For example,

in models with equal probabilities to be viable this happens if $p < 1/(2n)$. In the subcritical regime, the largest clusters of viable genotypes in the genotype space have size of order n and there are many of them. Typical members of a connected component are connected by a single path. This path is straight in the sense that along it, substitution in a locus can happen only once. The overall number of different species on a connected component has order n. The expected number of different species on a connected component within k viable substitution from any its member is of order k, i.e. is small.

The supercritical regime takes place if the proportion of viable genotypes is small but not extremely small. For example, in models with equal probabilities to be viable this happens if $p > 1/(2n)$. In the supercritical regime, there exists a cluster of viable genotypes (a "giant" component) that includes a positive proportion of all viable homozygotes. The "giant" component, which has size order of $2^n/n$, comes "near" every point of the genotype space. Typical members of the giant component are connected by many evolutionary paths which are not extremely windy. The number of different species on the giant component has at least order n^2. The expected number of different species on a connected component within k viable substitution from any its member is at least of order kn, i.e is very large. At the boundary of two regimes all properties of adaptive landscapes undergo dramatic changes, a physical analogy of which is a phase transition.

We have considered the most probable (within the present framework) scenario of biological evolution on holey landscapes assuming that it starts on a genotype from the largest connected component and proceeds along it by mutation and genetic drift. In this scenario, there is no need to cross any "adaptive valleys". Reproductive isolation between populations evolves as a side effect of accumulating different mutations. The rate of divergence is very fast: a few substitutions are sufficient to result in a new biological species (cf, Orr, 1995).

All of the above conclusions are qualitatively valid in three different models that we have considered, thereby suggesting their considerable generality.

RELATIONS TO SOME PREVIOUS APPROACHES/IDEAS

In this section we discuss relations of our results to some previous ideas and approaches.

Connected sets in the sequence space

Maynard Smith [1970; see also Conrad (1982) for discussion] argued that divergent protein evolution is impossible in historical time unless $fM > 1$, where M is the number of proteins which can be derived from a functionally useful protein and f is the fraction of these with an acceptable selective value. He stated that if $fM > 1$, then functional proteins form a continuous network in the protein space which can be traversed by unit mutational steps without passing through nonfunctional intermediates. Using our notation, with n diallelic loci, $M = 2n$, and with only two possible fitness values (0 or 1) $f = p$. Thus, Maynard Smith's condition corresponds to our condition $p > 1/(2n)$ for the supercritical regime under which there exists a giant connected component of viable genotypes which expands through the whole genotype space. Existence of very large connected sets of RNA sequences folding to the same secondary structures has been demonstrated in recent numerical works (Schuster *et al*., 1994; Huynen *et al*., 1996).

"*Extra-dimensional bypass*"

Conrad (1990) puts forward an idea of an "extra-dimensional bypass" on adaptive surfaces. According to Conrad an increase in the dimensionality of an adaptive landscape is expected to transform isolated peaks into saddle points that can be easily escaped resulting in continuing evolution. The increase of the dimensionality of the adaptive landscape might be a consequence of an increase in the size of genome. This idea is closely related to arguments used by Fisher in his critiques of Wright's presumption that selection would tend to confine populations to local peaks in an adaptive landscape and thus prevent them from finding higher peaks. Fisher (see Provine, 1986, pp. 274–275; Ridley, 1993, pp. 206–207) pointed out that as the number of dimensions in an adaptive topography increases, local peaks in lower dimensions tend to become saddle points in higher dimensions. In this case, according to Fisher, natural selection will be able to move the population to the global peak without any need for genetic drift.

Our results provide a formal justification of the idea of an "extra-dimensional bypass". Let us fix the number of loci and consider a population that belongs to a "small" connected component and, thus, has only limited possibilities to evolve. Assume also that in the genotype space there exists another "large" connected component, which, however, cannot be explored by the population. If the number of loci increases while p is kept constant, the two connected components will eventually belong to the same giant component with a positive probability. [A possible mechanism for increasing the number of loci is gene duplication. For a recent theoretical analysis see Walsh (1995)]. This follows from the fact that the

critical p value decreases as $n \to \infty$ and the systems moves to the supercritical regime where a positive proportion of all viable genotypes belong to the giant component. If p is small, any two viable genotypes will typically become connected by an evolutionary path when $n \approx 1/p$. That shows that increasing the number of loci would allow to explore the whole genotype space.

Relationships between species diversity and quality of the environment

As was mentioned above a key parameter of our model, probability p, can also be interpreted as a measure of environmental hostility: the smaller p is, the more difficult it is to survive. Our results indicate that the number of species in the largest connected component is a unimodal function of p, which achieves its maximum near the point of phase transition (see Fig. 3). Thus, the model predicts that "species diversity" (number of species) should be a hump-shaped function of the "quality" or "productivity" of the environment (measured by p) with maximum species diversity at intermediate values of quality of the environment (cf, Rosenzweig & Abramsky, 1993).

Evolution on "rugged" landscapes

The results presented here allow to get some additional information about uncorrelated "rugged" landscapes of Kauffman and Levin (1987). These fitness landscapes arise if genotype fitness, w, is a realization of a random variable having uniform distribution between zero and one. Assume that there is a rugged landscape. Let us introduce a threshold value, $w_c = 1 - p$, and construct a holey landscape in

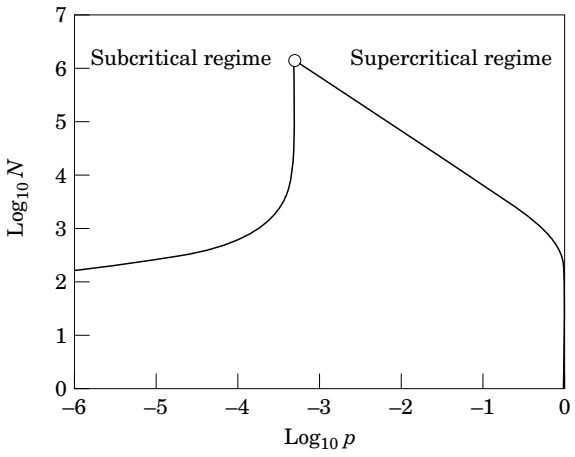

FIG. 3. Number of species in the largest connected component, N, as function of p on a log-log scale for $n = 1000$. The circle marks the point of phase transition at $p \approx 1/(2n)$.

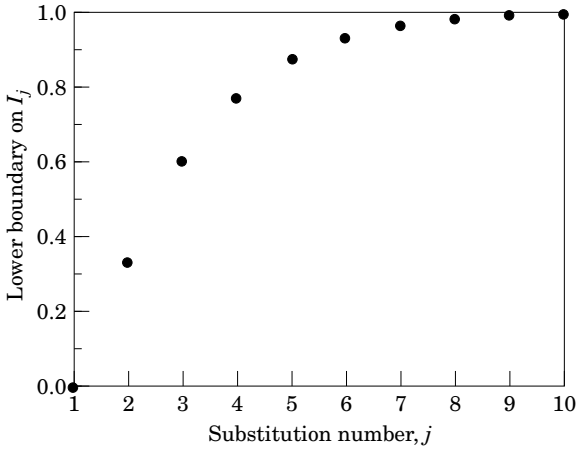

FIG. 4. The lower bound on the probability of speciation I_j after substitution number j.

that a genotype has fitness 1 if its fitness in the rugged landscape is larger then w_c, and fitness 0, if its fitness in the rugged landscape is smaller or equal to w_c. According to our results on holey landscapes if $p > 1/(2n)$, there exists a giant component of viable genotypes which extend throughout the whole genotype space. This giant component is generated by genotypes that have fitness at least $1 - p$ in the corresponding rugged landscape. That means that in the rugged landscapes there are very high "ridges" (with genotype fitnesses between $1 - p$ and 1) that continuously extend throughout the genotype space. In a similar way, if we choose $w_c = p$, it follows that the rugged landscapes have very deep "gorges" (with genotype fitnesses between 0 and p) that also continuously extend throughout the genotype space. Finally, one can choose two threshold values, w_{c1} and w_{c2} such that $w_{c1} - w_{c2} = p$, and construct a holey landscape in that a genotype has fitness 1 if its fitness in the corresponding rugged landscape is between w_{c1} and w_{c2}. Proceeding as before, one can show that the rugged landscape has "levels" with genotype fitnesses between w_{c1} and w_{c2} that again continuously extend throughout the genotype space.

A finite population subject to mutation is likely to be found on a fitness level determined by mutation-selection-random drift balance. Genotypes with fitnesses close to this level form a corresponding giant component. The population is prevented by selection from "slipping" off this component to genotypes with lower fitness and by mutation (and recombination) from "climbing" to genotypes with higher fitness. A population which has reached the giant component should be kept on it and further evolution should proceed in a quasi neutral fashion according to the properties of holey landscapes (cf, Woodcock &

Higgs, 1996). According to this scenario, microevolution and local adaptation can be viewed as climbing of the population towards the holey landscape, whereas macroevolution and speciation can be viewed as a movement of the population along the holey landscape.

We are grateful to Michael Conrad, Lev Ginzburg, Carole Hom, Bruce Walsh and a reviewer for helpful comments on the manuscript. JG was partially supported by the research grant J1–6157–0101–94 from the Republic of Slovenia's Ministry of Science and Technology. SG was partially supported by U.S. Public Health Service Grant R01 GM 32130 to Alan Hastings.

REFERENCES

Ajtai, M., Komlós J. & Szemerédi, E. (1982). Largest random component of a k-cube. *Combinatorica* **2**, 1–7.

Barton, N. H. (1989). Founder effect speciation. In: *Speciation and its Consequences* (Otte D. & Endler J. A., eds), pp. 229–256. MA: Sunderland.

Barton, N. H. & Bengtsson, B. O. (1986). The barrier to genetic exchange between hybridizing populations. *Heredity* **56**, 357–376.

Barton, N. H. & Rouhani, S. (1993). Adaptation and the shifting balance. *Genet. Res.* **61**, 57–74.

Bengtsson B. O. & Christiansen, F. B. (1983). A two-locus mutation selection model and some of its evolutionary implications. *Theor. Popul. Biol.* **24**, 59–77.

Bengtsson, B. O. (1985). The flow of genes through a genetic barrier. In: Evolution Essays in Honor of John Maynard Smith (Greenwood J. J., Harvey, P. H. & Slatkin, M., eds), pp. 31–42. Cambridge: Cambridge University Press.

Bollobás, B. (1985). *Random Graphs*. London: Academic Press.

Bollobás, B., Kohayakawa, Y. & Łuczak, T. (1992). The evolution of random subgraphs of the cube. *Random Structures and Algorithms* **3**, 55–90.

Bollobás, B., Kohayakawa, Y. & Łuczak, T. (1995). Connectivity properties of random subgraphs of the cube. *Random Structures and Algorithms* **6**, 221–230.

Bollobás, B. & Leader, I. (1990). Exact face-isoperimetric inequalities. *European J. Combinatorics* **11**, 335–340.

Bollobás, B. & Thomason, A. (1985). Random graphs of small order. In: "*Random Graphs* '83", pp. 47–97. Annals of Discrete Mathematics 28, North-Holland.

Burtin, Yu. D. (1977). On the probability of connectedness of a random subgraph of the n-cube. *Problemy Pered. Inf.* **13**, 90–95. (in Russian)

Cabot, E. L., Davis, A. W., Johnson, N. A. & Wu, C.-I. (1994). Genetics of reproductive isolation in the *Drosophila simulans* claude: complex epistasis underlying hybrid male sterility. *Genetics* **137**, 175–189.

Carson, H. L. (1968). The population flush and its genetic consequences. In: *Population Biology and Evolution* (Lewontin, R. C., ed.), pp. 123–137. Syracuse: Syracuse University Press, NY.

Conrad, M. (1982). Natural selection and the evolution of neutralism. *BioSystems* **15**, 83–85.

Conrad, M. (1990). The geometry of evolution. *BioSystems* **24**, 61–81.

Conway, J. & Sloane, N. (1988). *Sphere Packings, Lattices and Groups*. New York: Springer-Verlag.

Coyne, J. A., Barton, N. H. & Turelli, M. (1996). A critique of Sewall Wright's shifting balance theory of evolution. *Evolution* (submitted).

Dobzhansky, T. H. (1937). *Genetics and the Origin of Species*. New York: Columbia University Press.

Dyer, M. E., Frieze, A. M. & Foulds, L. R. (1987). On the strength of connectivity of random subgraphs of the n-cube. In: "*Random Graphs* '85", pp. 17–40. Annals of Discrete Mathematics 33, North–Holland.

Erdös, P. & Rényi, A. (1959). On random graphs. *Publ. Math. Debrecen* **6**, 290–297.

Erdös, P. & Spencer, J. (1979). Evolution of the n-cube. *Computers and Math. with Appls.* **5**, 33–40.

Gavrilets, S. (1996). On phase three of the shifting-balance theory. *Evolution* **50**, 1034–1041.

Gavrilets, S. & Hastings, A. (1995). Dynamics of polygenic variability under stabilizing selection, recombination, and drift. *Genet. Res.* **65**, 63–74.

Gavrilets, S. & Hastings, A. (1996). Founder effect speciation: a theoretical reassessment. *Amer. Nature* **147**, 466–491.

Ginzburg, L. R. & Braumann, C. A. (1980). Multilocus population genetics: relative importance of selection and recombination. *Theor. Popul. Biol.* **17**, 298–320.

Grimmett, G. (1989). *Percolation*. New York: Springer-Verlag.

Huynen, M. A., Stadler, P. F., & Fontana, W. (1996). Smoothness within ruggedness: the role of neutrality in adaptation. *Proc. Natl. Acad. Sci. U.S.A.* **93**, 397–401.

Karlin, S. & Carmelli, D. (1975). Numerical studies on two-locus selection models with general viabilities. *Theor. Popul. Biol.* **7**, 399–421.

Kauffman, S. A. & Levin, S. (1987). Towards a general theory of adaptive walks on rugged landscapes. *J. theor. Biol.* **128**, 11–45.

Kesten, H. *Asymptotics in high dimension for percolation.* Unpublished manuscript.

Lande, R. (1979). Effective deme size during long-term evolution estimated from rates of chromosomal rearrangements. *Evolution* **33**, 234–251.

Lande, R. (1985). The fixation of chromosomal rearrangements in a subdivided population with local extinction and recolonization. *Heredity* **54**, 323–332.

Lewontin, R. C., Ginzburg, L. R. & Tuljapurkar, S. D. (1978). Heterosis as an explanation for large amounts of genetic polymorphism. *Genetics* **88**, 149–170.

MacWilliams, F. & Sloane, N. (1977). *The Theory of Error–Correcting Codes*. North-Holland: Elsevier.

Maynard Smith, J. (1970). Natural selection and the concept of a protein space. *Nature* **225**, 563–564.

Mayr, E. (1942). *Systematics and the Origin of Species*. New York: Columbia University Press.

Mayr, E. (1954). Change of genetic environment and evolution. In: *Evolution as a Process* (Huxley J. S., Hardy, C. & Ford, E. B., eds), pp. 156–180. London: Allen and Unwin.

Mayr, E. (1963). *Animal Species and Evolution*. Cambridge: Harvard University Press, MA.

Nei, M., Maruyama, T. & Wu, C.-I. (1983). Models of evolution of reproductive isolation. *Genetics* **103**, 557–579.

Orr, H. A. (1995). The population genetics of speciation: the evolution of hybrid incompatibilities. *Genetics* **139**, 1803–1813.

Palmer, E. M. (1985). *Graphical Evolution*. New York: Wiley.

Provine, W. B. (1986). *Sewall Wright and Evolutionary Biology* Chicago and London: The University of Chicago Press.

Ridley, M. (1993). *Evolution* Boston: Blackwell Scientific Publications.

Rosenzweig, M. L. & Abramsky, Z. (1993). How are diversity and productivity related? In: *Species Diversity in Ecological Communities* (Ricklefs, R. E. & Schluter, D., eds), pp. 52–65. Chicago: The University of Chicago Press.

Rouhani, S. & Barton, N. H. (1993). Group selection and the shifting balance. *Genet. Res.* **61**, 127–135.

Schuster, P., Fontana, W., Stadler, P. F. & Hofacker, I. L. (1994). From sequences to shapes and back: a case study in RNA secondary structures. *Proc. Roy. Soc. Lond. B* **255**, 279–284.

Templeton, A. R. (1980). The theory of speciation via the founder principle. *Genetics* **94**, 1011–1038.

Turelli, M. & Ginzburg, L. (1983). Should individual fitness increase with heterozygosity? *Genetics* **104**, 191–209.

WAGNER A., WAGNER, G. P. & SIMILION, P. (1994). Epistatis can facilitate the evolution of reproductive isolation by peak shifts: a two-locus two-allele model. *Genetics* **138**, 533–545.
WALSH, J. B. (1995). How often do duplicated genes evolve new functions? *Genetics* **139**, 421–428.
WHITLOCK, M. C., PHILLIPS, P. C., MOORE, F. B.-G. & TONSOR, S. J. (1995). Multiple fitness peaks and epistatis. *Annu. Rev. Ecol. Syst.* **26**, 601–629.
WOODCOCK, G. & HIGGS, P. G. (1996). Population evolution on a multiplicative single-peak fitness landscape. *J. theor. Biol.* **179**, 61–73.
WRIGHT, S. (1931). Evolution in Mendelian populations. *Genetics* **16**, 97–159.
WRIGHT, S. (1980). Genic and organismic selection. *Evolution* **34**, 825–843.

APPENDIX

In this section, we very briefly sketch the proofs of the main results. Detailed proofs will appear elsewhere.

We start by remarking that the problem of determining N for general n and n_{min} arises in the theory of error correcting codes (MacWilliams & Sloane, 1977; Conway & Sloane, 1988). The results (1a) for $n_{2\min} = 2$ follow from simple computations:

$$N(n, 2) = \sum_{k \bmod 2 = 0} \binom{n}{k} = 2^{n-1}, \qquad \text{(A.1)}$$

$$3^n \pi(n, 2) = 2 \cdot \left(A + \sum_{k \bmod 2 = 1} \binom{n-1}{k} \right) - 1$$

$$= 3 \cdot 2^{n-1} - 1. \quad \text{(A.2)}$$

Then formulae (1b) for $n_{min} = 3$ follow from the following identity, valid for odd n_{min}: $N(n-1, n_{min}-1) = N(n, n_{min})$.

Result 1(a). This can be proved using results in (Bollobás *et al*., 1992).

Result 1(b). The upper bound is quite straightforward, because the probability of having k viable species in the *entire* B_n is at most

$$\binom{2^n}{k} p^k (1-p)^{k(k-1)/2},$$

which gives the desired bound. To prove the lower bound in the supercritical regime, one uses the ideas from Bollobás *et al*. (1992) and Bollobás & Thomason (1985). To give the indication how this works, fix a large number N of homozygotes and a small $p > 0$. How many species can we expect (ignore the connectivity)? Put the N genotypes in a row, and examine them one by one. The second species emerges after about e^p examinations (actually $1/(1-p)$, but p is small), and for the third, about e^{2p} more examinations are needed. The number of species k thus satisfies the approximate equation:

$$e^{kp} \approx N \Rightarrow k \approx \frac{\ln N}{p}.$$

This gives a relatively small number of genotypes versus the size of the hypercube, so it is relatively easy to connect a positive proportion of them together.

Finally, a standard branching process comparison handles the subcritical regime.

Result 1(c). The arguments here are variations on those found in the random graph literature, see Bollobás (1985), Bollobás *et al*. (1992, 1995), Dyer *et al*. (1987), and Palmer (1985).

Result 1(d). These statements are results of a combination of combinatorial arguments and those from references above.

Result 2(a). This is proved by standard methods, as found in Bollobás *et al*. (1995).

Result 2(b). The upper bound follows from the fact that for any set $G \subset Q_n$ of k homozygotes, the smallest possible number of heterozygotes obtained by mating genotypes in G is $k^{3/2}$ [see Bollobás & Leader (1990) for a similar result]. For the lower bound, one estimates the probability of the event that no heterozygotes are viable on the sub-cube $Q_{\lfloor \log_2(k/p) \rfloor}$, while all homozygotes are connected to the giant component.

The following non-rigorous, but convincing, argument shows that N_1 should be of order n/p, for n large and $p \geqslant 1/n$ small.

Imagine that $Q_n \subset Q_{n+1}$, by adding a 0 at the end of every genotype in Q_n. If we have k species in Q_n, then the expected number of all genotypes in Q_{n+1}/Q_n which produce inviable genotype by mating with every one of k species is

$$2^n \cdot (1-p)^k \approx e^{n \log 2 - pk}.$$

Assume $k = cn/p$, for a constant c. If $c < \log 2$, the number is exponentially large and, presumably, it is easy to select a subset of size $1/p$ consisting of different species from this large set. On the other hand, if $c > \log 2$, then with probability close to 1 there is not even one new species in the entire Q_{n+1}/Q_n.

Result 3(a). The supercritical part (8b) follows directly from Ajtai *et al*. (1982) and Bollobás *et al*. (1992). To prove the subcritical part (8a), we start by observing that the probability that, say, $(0, \ldots, 0)$ survives on a self-avoiding path of m steps is bounded

above by

$$p^p \sum_{\text{paths of length m}} a^{\text{number of 1's on the path}}.$$

A slightly involved combinatorial argument can be used to get an upper bound on this expression.

Result 3(*b*). The first step is to prove that, with overwhelming probability, no genotype with $c\sqrt{n}$ or more heterozygotic loci is in L_1, given that $c > 2/\log(1/a)$. Then one can use bounds from the theory of error-correcting codes (MacWilliams & Sloane, 1977; Conway & Sloane, 1988) to get a set S_1 of, say, $0.1n$-separated homozygotes from L_1 with size $|S_1| \geqslant e^{0.3n}$, which are with high probability all different species.