Evolution in a flat fitness landscape

Bernard Derrida, Luca Peliti

Bulletin of Mathematical Biology Vol. 53, No. 3, pp. 355–382, 1991.
Printed in Great Britain.

0092–8240/91$3.00 + 0.00
Pergamon Press plc

EVOLUTION IN A FLAT FITNESS LANDSCAPE

■ BERNARD DERRIDA*† and LUCA PELITI‡§
*School of Mathematics,
Institute for Advanced Studies,
Princeton, NJ 08540, U.S.A.
‡Groupe de Physico-Chimie Théorique,
URA 1382, E.S.P.C.I.,
10, rue Vauquelin,
F-75231 Paris Cedex 05, France

A simple model of a population of asexually reproducing individuals, evolving in a flat fitness landscape, is defined. It is shown that the model is equivalent to a dynamical system with stochastic dynamics, the Annealed Random Map Model. Thus, it is possible to solve exactly for the genealogy statistics and for the genetic variability of the population. Fluctuations of quantities, like the average relatedness and the variability, which also take place in the limit of an infinitely large population, are computed.

1. Introduction. Several models of Darwinian evolution at the molecular level have recently been introduced. They have different structures, according to which features are more of interest to the investigators. In adaptive walks (see, e.g. Kauffman and Levin, 1987; Kauffman, 1989; Macken and Perelson, 1989) evolution is modelled as a stepwise optimization process: the genotype of a species is mutated at random, and the new genotype is adopted if it leads to a higher fitness. This approach does not take explicit account of the genetic variability of the population. The quasi-species model (see Eigen *et al.*, 1988 for a recent review) represents evolution by deterministic equations modelled on those of reaction kinetics. The state of a population is described by the relative frequency of each molecular species. The approach neglects fluctuations in these quantities: this is in principle warranted if the number of possible molecular species is much smaller than the total number of molecules. Although this is the usual situation in reaction kinetics, it appears of doubtful validity in this context, since the number of possible molecular species (e.g. of possible nucleotide sequences in polynucleotide dynamics) is staggering.

† Permanent address: Service de Physique Théorique de Saclay, Institut de Recherche Fondamentale, F-91191 Gif-Sur-Yvette Cedex, France.

§ Permanent address: Dipartimento di Scienze Fisiche and Unità INFM, Università di Napoli, Mostra d'Oltremare, Pad. 19, I-80125 Napoli, Italy.

Nevertheless, the main results of quasi-species theory hold if fitness optima are well pronounced and sparse. However, a number of features related to finite population size do require a stochastic treatment (Ebeling and Feistel, 1977; Swetina and Schuster, 1982; Schuster and Sigmund, 1985; Schuster and Swetina, 1988). A very instructive phenomenon is "Muller's ratchet", in which a population moves away from an adaptation optimum because of fluctuations (Nowak and Schuster, 1989).

Stochastic models have been mainly treated by computer simulations (Fontana and Schuster, 1987; Fontana *et al.*, 1989; Amitrano *et al.*, 1988; 1989), although some analytical results have been obtained (Demetrius *et al.*, 1985). Some important properties have thus been highlighted. Of course, the case in which a stochastic treatment is most needed is that of a neutral or almost neutral evolution (Section 8.1 of Kimura, 1983). In this case, the deterministic quasi-species equations can be trivially solved, yielding the prediction that, at equilibrium, the frequency of any allowed molecular species is as large as that of any other. But it is obvious that the finite size of the population implies that, at any given time, only a few molecular species can possibly be represented.

In fact, the description of a stochastic evolutionary model requires the introduction of two different kinds of averages. One might take the average of relevant properties over the whole molecular population existing at any given time: this is the population average. Quantities like the consensus sequence and the genetic variability are defined in this way. However, these quantities fluctuate *even for very large populations*. One can thus envisage taking their average over a very long time stretch: this is a time average, which, if some sort of ergodic property is assumed, may be represented by the average over all possible realizations of the stochastic process. One of the weaknesses of the standard quasi-species model lies, in our opinion, in not differentiating between these two kinds of averages.

The presence of two kinds of averages is commonplace in the theory of disordered systems, and in particular of spin glasses (a useful review and collection of papers is contained in Mézard *et al.*, 1987). The population average is analogous to the thermal average, and the long time or process average to the average over disorder. The stochastic nature of thermal averages goes under the name of "lack of self-averaging".

We show in this paper how these properties appear in the explicit solution of a very simple model of a population evolving in a flat fitness landscape. The model is inspired by those of Fontana and Schuster (1987), Fontana *et al.* (1989), Amitrano *et al.* (1988; 1989) and Zhang *et al.* (1989).

Although the consideration of a flat fitness landscape may appear academic, we think that our results are relevant for two reasons: on the one hand, the explicit knowledge of the behavior of a population evolving in the absence of Natural Selection makes it easier to identify, by contrast, the features which are

due to Natural Selection itself; on the other hand, we shall argue in the final discussion that our results about genealogy statistics also hold for a model of a population evolving in a rugged fitness landscape (Kauffman and Levin, 1987; Kauffman, 1989; Macken and Perelson, 1989), at least in the infinite genome size limit (p. 236 of Kimura, 1983).

We shall use the following strategy: the problem of the distribution in genome space of the population is conceptually split into genealogy statistics on the one hand, and drift in genome space of a given lineage on the other. If two individuals share a more or less recent common ancestor, they will have a greater or lesser genetic similarity. Now, we are able to compute the genealogy statistics of our model, by exploiting its equivalence with a dynamical system with stochastic dynamics, the Annealed Random Map model, introduced by Derrida and Bessis (1988). A new link between models in evolution theory and dynamical systems is thus established (different from those discussed by Hofbauer and Sigmund, 1988). The drift in genetic space then appears as a simple random walk problem, whose solution allows us to predict explicitly the distribution of genetic similarity and the statistics of the genetic variability of the population.

The model is defined in Section 2. The statistics of genealogies is derived in Section 3, and some of its consequences are explored in Section 4. Section 5 contains a discussion of the genetic variability of the population, and Section 6 discusses its corresponding fluctuations. Section 7 contains a discussion of the genetic drift. A general discussion closes the paper, while two technical results, obtained by Derrida and Bessis (1988), are rederived in the Appendices for completeness.

2. Model. We consider a population Ω, made up of a fixed number, M, of individuals reproducing asexually, whose genome is characterized by N binary units $S_i^\alpha = \pm 1$, $\alpha = 1, 2, \ldots, M$; $i = 1, 2, \ldots, N$. At each generation t, all individuals are removed, and a new generation is formed by offsprings of the previous individuals. To each individual $\alpha \in \Omega$ is assigned a parent $G_t(\alpha) \in \Omega$. We stipulate that $G_t(\alpha)$ is chosen independently and with uniform probability in Ω for each individual α and at each generation t. The fitness landscape is flat, in the sense that all genotypes have equal chances of leaving behind offspring.

Therefore, for finite values of M, the probability p_m that an individual leaves behind m offspring is a binomial with success probability equal to $1/M$. If $M \gg 1$, this probability becomes a Poisson distribution of mean 1:

$$p_m = \frac{e^{-1}}{m!}. \tag{1}$$

The genome of each offspring would be identical to that of its parent, were it not for the rare occurrence of mutations. We consider here only point

mutations. This means that, at each generation t, the genome $\mathbf{S}^\alpha = (S_i^\alpha)$ of the individual α is identical to that of its parent $G_t(\alpha)$ at the previous generation $t-1$, except for mutations, which occur with probability $\mu\, dt$ during each time interval dt. Therefore:

$$S_i^\alpha(t) = \begin{cases} S_i^{G_t(\alpha)}(t-1), & \text{with probability } \frac{1}{2}(1+e^{-2\mu}), \\ -S_i^{G_t(\alpha)}(t-1), & \text{with probability } \frac{1}{2}(1-e^{-2\mu}). \end{cases} \tag{2}$$

One might also consider more general models, where the probability p_m is not given by equation (1). It turns out that our results apply also to these cases up to a simple rescaling of time, provided that all individuals are equal with respect to their chances of reproducing. This point will be discussed in the next section.

We now show that the genealogy statistics of our model can be represented by the Annealed Random Map (ARM) model, introduced and solved by Derrida and Bessis (1988). In the rest of this section—and in the following one—we forget about the genome structure of the individuals and consider only their labels $\alpha = 1, 2, \ldots, M$.

The ARM model is defined as follows. One considers a phase space Ω, made up of M points. At each time step t, the dynamics is defined by a random mapping G_t: $\Omega \to \Omega$ of Ω into itself. For each point $\alpha \in \Omega$ one chooses independently and with uniform probability in Ω its image $G_t(\alpha)$. Mappings at different times t are independent.

As time goes on, the images of different points may be mapped on the same point, i.e. trajectories may merge. Let us follow the fate, after t time steps, of the images of the different points in Ω for a given sequence of mappings G_t. The phase space Ω splits up in a certain number of "valleys" such that, if two points α and β belong to the same valley, they are mapped on the same image after t time steps. The valleys depend on t (hence they may be aptly called "t-valleys") and on the particular sequence of mappings (G_t).

Let us now consider our population, a long time after the beginning. We can now investigate the genealogy of a given individual by tracing *backwards* its parent and ancestors: as we move back one generation, the parent is identified by applying the random mapping G_t once, and the ancestors by applying the corresponding mapping to the parent, and so on. Merging of the images of two different points corresponds to the *splitting* of two lineages in genealogy statistics. By the same token, the fact that two points, α and β, belong to the same "t-valley" in the ARM model, is equivalent in the population model to the fact that two individuals had a common ancestor t generations ago.

Building on this analogy we can now restate, in the language of genealogy statistics, some of the results obtained by Derrida and Bessis (1988) for the ARM model.

3. Genealogies. Given n individuals in the population, let us denote by $w_n(t)$ the probability that their ancestors, t generations ago, are all different. This quantity is easily calculated. If k individuals have already been assigned each a different parent, the next one will be assigned another different one with probability $(1-k/M)$. Therefore, the probability x_k that k individuals have k different parents is given by:

$$x_k=\left(1-\frac{1}{M}\right)\left(1-\frac{2}{M}\right)\ldots\left(1-\frac{k-1}{M}\right). \tag{3}$$

As a consequence, one has:

$$w_n(t+1)=\left(1-\frac{1}{M}\right)\left(1-\frac{2}{M}\right)\ldots\left(1-\frac{n-1}{M}\right)w_n(t). \tag{4}$$

By expanding this relation up to first order in $1/M$ we obtain:

$$w_n(t+1)=w_n(t)\left[1-\frac{n(n-1)}{2}\frac{1}{M}\right]. \tag{5}$$

This allows us to derive the expression of $w_n(t)$:

$$w_n(t)=\exp\left[-\frac{n(n-1)}{2}\frac{t}{M}\right]. \tag{6}$$

We see that the relevant time scale is proportional to M. We therefore introduce the rescaled time variable:

$$\tau=\frac{t}{M}, \tag{7}$$

and express all time dependences in terms of τ.

It is now easy to see that, in a more general reproduction scheme characterized by a probability p_m of leaving behind m offspring different from equation (1), equation (6) must be replaced by:

$$w_n(t+1)=w_n(t)\left[1-\frac{n(n-1)}{2}\frac{\kappa}{M}\right], \tag{8}$$

with κ defined by:

$$\kappa=\sum_{m=2}^{\infty} m(m-1)p_m=[m^2]_{\text{av}}-1, \tag{9}$$

where the average $[\ \ ]_{\text{av}}$ is taken with respect to the probability distribution

p_m, and we have exploited the fact that, since we impose the population size M to be constant, one has $[m]_{\text{av}}=1$. Our results therefore apply also to this case, provided we define the rescaled time variable τ as follows:

$$\tau=\kappa\frac{t}{M}. \tag{10}$$

In particular, in the scheme defined by Amitrano *et al.* (1989) and Zhang *et al.* (1989) a fraction δ of the individuals is suppressed outright, and the remaining ones have one or more offspring, as they are allowed to reproduce to fill in the gaps of the population. One has in this case:

$$p_0=\delta;\quad p_m=\frac{1-\delta}{(m-1)!}\left(\frac{\delta}{1-\delta}\right)^{m-1}\exp\left(-\frac{\delta}{1-\delta}\right),\ m>0. \tag{11}$$

One then has:

$$\kappa=\frac{\delta(2-\delta)}{1-\delta}. \tag{12}$$

We shall call this scheme the ZSP model.

For the case of two individuals, we have from equation (6):

$$w_2(\tau)=\mathrm{e}^{-\tau}, \tag{13}$$

which implies that the probability $X_2(\tau)$ that two individuals had a common ancestor τM generations ago is given by:

$$X_2(\tau)=1-\mathrm{e}^{-\tau}. \tag{14}$$

The probability density $p(\tau)$ that the last common ancestor of the two individuals existed between τM and $(\tau+\mathrm{d}\tau)M$ generations ago is given by:

$$p(\tau)=\frac{\mathrm{d}X_2(\tau)}{\mathrm{d}\tau}=\mathrm{e}^{-\tau}. \tag{15}$$

This quantity may be generalized to the probability of a given genealogy, which may be represented by a set of genealogical trees such as the one shown in Fig. 1. We denote unbranched lineages by straight lines, and we identify by the time labels $0\leqslant\tau_n\leqslant\tau_{n-1}\ldots\leqslant\tau_k$, the epochs at which the different lineages branched. The label l identifies the number of different lineages in the generations immediately following τ_l. These variables are always rescaled numbers of generations. Then, the probability $\Pi_n(\tau_n,\tau_{n-1},\ldots,\tau_k)$ of a given genealogy depends only on the number n of individuals one considers, on the

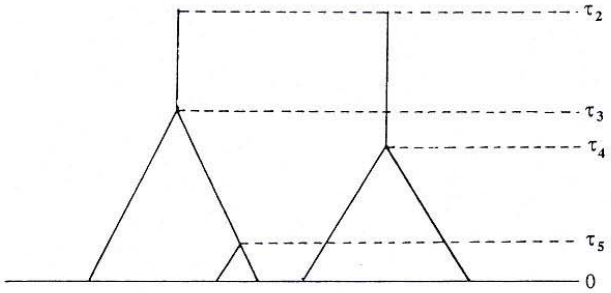

Figure 1. A genealogy with $n=5$, $k=2$. The variable τ increases as we reckon back the genealogy. The probability of this genealogy is given by $e^{-10\tau_5}e^{6(\tau_5-\tau_4)}e^{3(\tau_4-\tau_3)}e^{(\tau_3-\tau_2)}$.

number k of different ancestors, and on the branching times $\tau_n, \tau_{n-1}, \ldots, \tau_k$. One has:

$$\Pi_n(\tau_n, \tau_{n-1}, \ldots, \tau_k)=\prod_{l=k}^{n}\exp\left[-\frac{l(l-1)}{2}(\tau_l-\tau_{l+1})\right], \tag{16}$$

where we have defined $\tau_{n+1}=0$.

By the same token, $X_2(\tau)$ may be generalized to the set of probabilities $X_{n_1 n_2 \ldots n_k}(\tau)$ that, given $n=n_1+n_2+\cdots+n_k$ individuals, the first n_1 ones had, τM generations ago, one common ancestor, the second n_2 ones a different one, ... and the last n_k ones another different one. One finds in Derrida and Bessis (1988) the following expression for this quantity:

$$X_{n_1 n_2 \ldots n_k}(\tau)=n_1!n_2!\ldots n_k!(-1)^k \times\sum_{l=k}^{n}(-1)^l\frac{(2l-1)\,(n-k)!(k+l-2)!}{(n+l-1)!(n-l)!(l-k)!}\exp\left[-\frac{l(l-1)}{2}\tau\right]. \tag{17}$$

For reference, the derivation of this result is reported in Appendix 1.

It is important to keep in mind that two different kinds of average are involved in the model. Let us consider a particular history of the model, i.e. a particular time sequence of mappings (G_t) form each individual to its parent. At any given time, we may take the average of any quantity involving one or few individuals over the whole population Ω: we refer to it as the *population average* and denote it by angular brackets: $\langle\ \rangle$. However, such averages may fluctuate, even for an infinitely large population, according to the particular mapping sequence (G_t) which has taken place. One can therefore consider the average of these quantities, taken over all possible realizations of the reproduction process, i.e. over all possible sequences of mappings (G_t). We call it the *process average* and denote it by a bar: $\overline{\ }$.

There is a loose but instructive analogy between these averages and those appearing in the theory of disordered systems. The $M \to \infty$ limit in a population with a given history is analogous to the thermodynamical limit with a given realization of the quenched disorder. The population average is therefore analogous to the *thermal* average in disordered systems. The residual randomness makes it necessary to perform a further average over the disorder, which corresponds therefore to the process average. It is a special feature of our model, however, that the process average can be obtained by averaging over the temporal unfolding of the process for a sufficiently long time stretch, since the time sequence (G_t) of mappings belonging to different time intervals are independent.

For example, given the two individuals α and β, we may denote by $\tau_{\alpha\beta}$ their *relatedness*, i.e. the rescaled number of generations which we have to reckon back to find their last common ancestor. We then define $\vartheta_{\alpha\beta}(\tau)$ by:

$$\vartheta_{\alpha\beta}(\tau) = \theta(\tau - \tau_{\alpha\beta}), \tag{18}$$

where $\theta(\tau)$ is Heaviside's unit step function. We can thus define the quantity:

$$Y(\tau) = \langle \vartheta(\tau) \rangle = \left[\binom{M}{2} \right]^{-1} \sum_{(\alpha,\beta)} \vartheta_{\alpha\beta}(\tau), \tag{19}$$

where the sum is over all pairs of individuals in the population. $Y(\tau)$ is the probability that two individuals, chosen at random in the population present at a given epoch, had an ancestor in common τM generations before. One can show that the quantity $Y(\tau)$ is a random variable, which depends on the history of the population and fluctuates even when its size is infinitely large. The average of $Y(\tau)$ is given by:

$$\overline{Y(\tau)} = X_2(\tau) = 1 - e^{-\tau}. \tag{20}$$

To convince us that $Y(\tau)$ is indeed a random variable, let us consider:

$$\overline{Y^2(\tau)} = \overline{\langle \vartheta(\tau) \rangle^2} = \left[\binom{M}{2} \right]^{-2} \overline{\sum_{(\alpha,\beta)} \vartheta_{\alpha\beta}(\tau) \sum_{(\gamma,\delta)} \vartheta_{\gamma\delta}(\tau)}. \tag{21}$$

This quantity is the probability that α and β shared an ancestor τM generations ago, and γ and δ another, not necessarily different from the first. We have therefore, from equation (17):

$$\begin{aligned} \overline{Y^2(\tau)} &= X_4(\tau) + X_{2,2}(\tau) \\ &= 1 - \tfrac{8}{5} e^{-\tau} + \tfrac{2}{3} e^{-3\tau} - \tfrac{1}{15} e^{-6\tau}, \end{aligned} \tag{22}$$

which is larger than $\overline{Y(\tau)}^2$ for $\tau \neq 0, \infty$.

4. Families. We say that two individuals, α and β, belong to the same τ-family if their last common ancestor existed less than τM generations ago. The number $F(\tau)$ of τ-families and the number M_l of their members are fluctuating variables, whose distribution can be easily obtained.

Let us denote by $\Phi(\tau)$ the average of $F(\tau)$:

$$\Phi(\tau)=\overline{F(\tau)}. \tag{23}$$

A simple mean field approximation (where $\overline{F^2}$ is replaced by $\bar{F}^2=\Phi^2$) would give:

$$\frac{d\Phi_{mf}}{d\tau}=-\frac{\Phi_{mf}(\Phi_{mf}-1)}{2}, \tag{24}$$

and therefore:

$$\Phi_{mf}(\tau)=(1-e^{-\tau/2})^{-1}. \tag{25}$$

Derrida and Bessis (1988) have calculated the probability $Z_k(\tau)$ that there are exactly k families. The result reads, in the $M\to\infty$ limit:

$$Z_k(\tau)=\frac{1}{k!(k-1)!}\sum_{l=k}^{\infty}(-1)^{k+l}(2l-1)\frac{(k+l-2)!}{(l-k)!}\exp[-\tfrac{1}{2}l(l-1)\tau]. \tag{26}$$

For reference, this result is derived in Appendix 2. One obtains therefore the exact expression:

$$\Phi(\tau)=\sum_{k=1}^{\infty}kZ_k(\tau)=\sum_{l=1}^{\infty}(2l-1)\exp\left[-\frac{l(l-1)}{2}\tau\right]. \tag{27}$$

This expression agrees with the mean field one [equation (25)] when $\tau\ll 1$, yielding:

$$\Phi(\tau)\simeq\frac{2}{\tau}, \tag{28}$$

as already obtained by Zhang *et al.* (1989). However, one finds deviations in the fluctuation-dominated range $\tau\geqslant 1$, where $F(\tau)\simeq 1$. It is also possible to calculate the mean square deviation $\Delta F(\tau)$ defined by $\Delta F^2(\tau)=\overline{F^2(\tau)}-\overline{F(\tau)}^2$. The expression of $\overline{F^2(\tau)}$ is given by equation (A2.8). We show in Fig. 2a the average of $F(\tau)$ over 1000 simulation runs, along with the predictions (25) and (27), and in Fig. 2b the simulation results for $\Delta F(\tau)/\overline{F(\tau)}$, compared with the theoretical prediction derived from equation (A2.8).

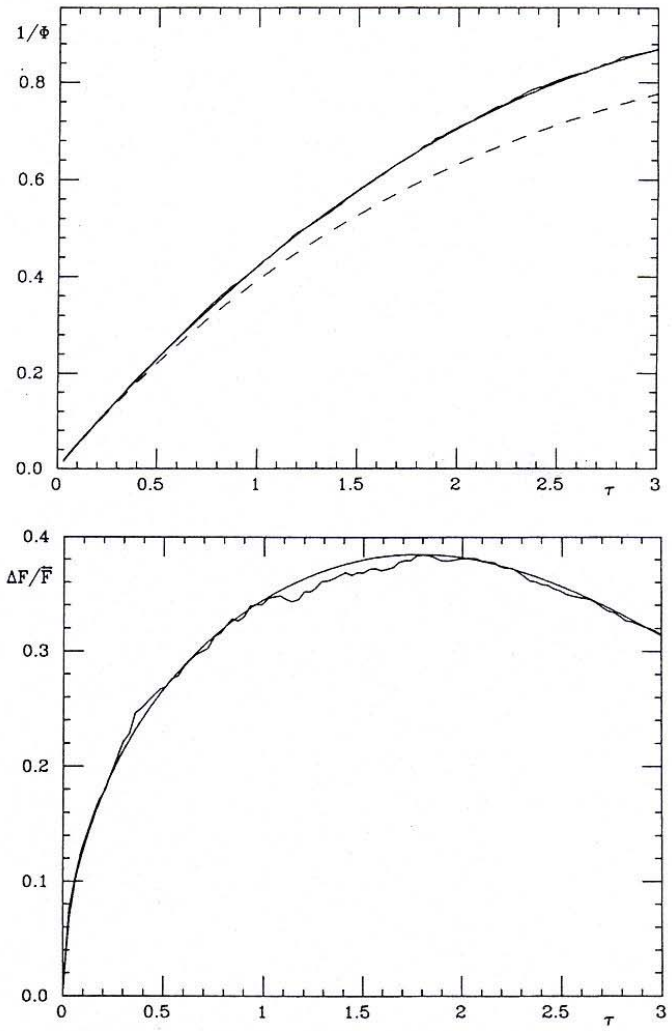

Figure 2. (a) Average number $\Phi(\tau)=\overline{F(\tau)}$ of τ-families *vs* τ for $M=500$. Solid line: equation (27). Broken line: mean field approximation equation (25). Irregular line: simulation. (b) Reduced mean square deviation $\Delta F(\tau)/\overline{F(\tau)}$, with $\Delta F^2=\overline{F^2}-\bar{F}^2$. Solid line: $\overline{F^2(\tau)}$ given by equation (A2.8). Irregular line: simulation. Simulation averages are taken over 1 000 runs, with $M=500$.

The distribution of sizes of τ-families has also been calculated by Derrida and Bessis (1988). The probability density $f(W_1, W_2, \ldots, W_k)$ that, if one chooses k families among all possible ones, the first one has $M_1=W_1M$ members, the second one $M_2=W_2M$, . . . and the last one has $M_k=W_kM$ members (where

the weights W_l satisfy $0 \leqslant W_l \leqslant 1$, and $\sum_l W_l \leqslant 1$), can be obtained from the requirement that:

$$X_{n_1 n_2 \ldots n_k}(\tau) = \int_0^1 \mathrm{d}W_1 \int_0^1 \mathrm{d}W_2 \ldots \int_0^1 \mathrm{d}W_k W_1^{n_1} W_2^{n_2} \ldots W_k^{n_k} f(W_1, W_2, \ldots, W_k). \tag{29}$$

The result is:

$$f(W_1, W_2, \ldots, W_k) = \sum_{l=k}^{\infty} \frac{l!(l-1)!}{(l-k)!} Z_l(\tau) \int_0^1 \mathrm{d}Y_1 \int_0^1 \mathrm{d}Y_2 \ldots \int_0^1 \mathrm{d}Y_l \times \delta(1 - Y_1 - Y_2 - \cdots - Y_l) \, \delta(W_1 - Y_1)\, \delta(W_2 - Y_2) \ldots \delta(W_k - Y_k). \tag{30}$$

This expression implies in fact that all possible ways of dividing the total population size M into the different family sizes M_l have the same probability, if the number k of families are given. In fact, if two τ-families merge into one at a given generation, this happens because the two ancestors of the first ones share a common parent, what takes place with a probability which is totally independent of their sizes.

It is now easy to calculate the probability distribution $\pi(Y)$ of $Y(\tau)$, defined in equation (19). Indeed, since W_l is the probability that a given individual belongs to the l^{th} τ-family, the probability $Y(\tau)$ that two individuals belong to the same τ-family may be expressed in terms of the weights W_l as follows:

$$Y(\tau) = \sum_l W_l^2, \tag{31}$$

where the sum runs over all τ-families. One gets (Derrida and Bessis, 1988):

$$\pi(Y) = Z_1(\tau)\, \delta(Y-1) + \sum_{k=2}^{\infty} (k-1)! Z_k(\tau) \int_0^1 \mathrm{d}W_1 \int_0^1 \mathrm{d}W_2 \ldots \int_0^1 \mathrm{d}W_k \times \delta(1 - W_1 - W_2 - \cdots - W_k)\, \delta(Y - W_1^2 - W_2^2 - \cdots - W_k^2). \tag{32}$$

The form of $\pi(Y)$ is qualitatively similar to that appearing in other disordered systems, such as spin-glasses (Derrida and Flyvbjerg, 1987; Mézard *et al.*, 1987).

5. Genetic Structure. Genealogies of reproducing molecules or bacteria are not accessible to experiment. It is however possible in principle to measure the genetic variability of a population, e.g. the probability that a given unit i is in a

state different from the most common one. In our model, a convenient measure of the variability is provided by the statistics of the overlap $q^{\alpha\beta}$ between two individuals α and β, defined by:

$$q^{\alpha\beta} = \frac{1}{N} \sum_{i=1}^{N} S_i^{\alpha} S_i^{\beta}. \tag{33}$$

Its population average $\langle q \rangle$ is given by:

$$\langle q \rangle = \left[\binom{M}{2} \right]^{-1} \sum_{(\alpha,\beta)} q^{\alpha\beta}, \tag{34}$$

where the sum runs over all distinct pairs of individuals in the population. When $\langle q \rangle \simeq 1$, the genetic variability of the population is small, and we have therefore a well defined quasi-species (Eigen *et al.*, 1988). On the other hand, when $\langle q \rangle \ll 1$, the population is widely spread in genome space. An even simpler quantity to compute is Q, defined by:

$$Q = \frac{1}{N} \sum_{i=1}^{N} \langle S_i^{\alpha} \rangle^2. \tag{35}$$

One has:

$$Q = \langle q \rangle + \frac{1}{M} (1 - \langle q \rangle) \simeq \langle q \rangle. \tag{36}$$

The probability distribution of the overlap $q^{\alpha\beta}$ can be explicitly calculated in our model.

Given two individuals, α and β, their Hamming distance $v^{\alpha\beta}$ is the number of genome units whose state is different in the two individuals:

$$v^{\alpha\beta} = \frac{1}{4} \sum_{i=1}^{N} (S_i^{\alpha} - S_i^{\beta})^2. \tag{37}$$

One has of course:

$$q^{\alpha\beta} = 1 - 2 \frac{v^{\alpha\beta}}{N}. \tag{38}$$

Let us now consider two individuals whose relatedness is equal to τ, i.e. whose last common ancestor existed τM generations ago. Their Hamming distance can be calculated as follows. During a time τM, the genome of the ancestor of each individual underwent a random walk in the space of the 2^N possible genomes, as mutations accumulated. The problem reduces therefore

to the study of a random walk in genome space, where each unit has a probability $\mu\, dt$ of flipping during each time interval dt.

Let us denote by $\phi_v(t)$ the probability that the total Hamming distance travelled in a walk lasting t generations, is equal to v. This probability $\phi_v(t)$ obeys the following equation:

$$\frac{d\phi_v(t)}{dt} = \mu[(v+1)\phi_{v+1}(t) + (N-v+1)\phi_{v-1}(t) - N\phi_v(t)]. \tag{39}$$

The solution of this equation reads:

$$\phi_v(t) = \frac{N!}{2^N v!(N-v)!}(1-e^{-2\mu t})^v(1+e^{-2\mu t})^{N-v}. \tag{40}$$

In particular the average Hamming distance $[v]_{av}(t)$ is given by:

$$[v]_{av}(t) = \sum_v v\phi_v(t) = \frac{N}{2}(1-e^{-2\mu t}). \tag{41}$$

We can now obtain the probability P_v that the Hamming distance between any two individuals is equal to v. If the last common ancestor of two individuals existed τM generations ago, they are separated by a random walk lasting $2\tau M$ generations. We have therefore:

$$P_v = \int_0^\infty d\tau p(\tau)\phi_v(2\tau M), \tag{42}$$

where $p(\tau)\, d\tau$ is the probability, given by equation (15), that the relatedness $\tau_{\alpha\beta}$ of the two individuals satisfies $\tau < \tau_{\alpha\beta} \leqslant \tau + d\tau$. The result reads:

$$P_v = \frac{\lambda\, N!}{2^N(N-v)!} \sum_{l=0}^{N-v} \binom{N-v}{l} \frac{\Gamma(\lambda+l)}{\Gamma(\lambda+l+v+1)}. \tag{43}$$

Here Γ is Euler's gamma function, and we have introduced the parameter

$$\lambda = \frac{1}{4\mu M}. \tag{44}$$

In the more general schemes one has:

$$\lambda = \frac{\kappa}{4\mu M}, \tag{45}$$

with κ given by equation (10). By use of equations (38) and (41) we also obtain:

$$\bar{Q}=\int_0^\infty d\tau e^{-\tau}e^{-4\mu M\tau}=\frac{\lambda}{\lambda+1}. \tag{46}$$

Equations (43) and (46) could in fact be derived in a different way by considering our model as the collection of N loci, each with two selectively equivalent alleles, and with mutation rate μ, evolving in a population of fixed size M. This is the $N>1$ generalization of the Wright model (Wright, 1937; see also Kimura and Crow, 1964; Stewart, 1976; p. 203 of Kimura, 1983). The same method allows us to treat a population evolving in a nontrivial (but highly correlated) fitness landscape, where each locus contributes independently to the death probability. However, in this way one would not recover the results for the genealogies derived in Section 3, and the calculation of fluctuations would be slightly harder.

The results simplify in the limit $N\gg 1$. In this case $\phi_v(t)$, given by equation (40), becomes approximately a delta function, centered on $[v]_{av}(t)$. We therefore have:

$$P_v=\int_0^\infty d\tau p(\tau)\,\delta(v-[v]_{av}(2\tau M))=\frac{2\lambda}{N}\left(1-\frac{2v}{N}\right)^{\lambda-1}. \tag{47}$$

By use of equation (38), we may convert this expression into the probability density $P(q)$ of the overlap q, obtaining:

$$P(q)=\begin{cases}\lambda q^{\lambda-1}, & \text{if } 0<q\leqslant 1,\\ 0, & \text{otherwise.}\end{cases} \tag{48}$$

We see that this simple model of evolution in a flat landscape presents a broad distribution $P(q)$ of the overlap q, of the same nature as the $P(q)$ predicted by the Parisi theory of spin glasses (Mézard *et al.*, 1987).

Figure 3 shows a plot of $\bar{Q}$ *vs* λ for our model (diamonds) and for the ZSP model (crosses). The points are averages over simulations, lasting 10 000 generations, for $M=200$ and $N=20$ and 10 respectively. The full line corresponds to equation (16). In spite of the smallness of the values of N and M, the agreement appears satisfactory.

Figures 4 and 5 show the histograms of the Hamming distance v, predicted by equation (43), compared with the results of the simulations of our model and of the ZSP model respectively. In both figures, the diamonds represent the theory and the histograms represent the outcome of the simulation, averaged over 10 000 generations. We remark that P_v has a peak away from $v=0$ for $\lambda<1$; the peak moves to $v=0$, becoming very flat at $\lambda=1$, and stays there, becoming sharper and sharper, as λ increases further.

It is also easy to generalize these results to the case of the diffusion–reproduction processes in real space considered by Zhang *et al.* (1989).

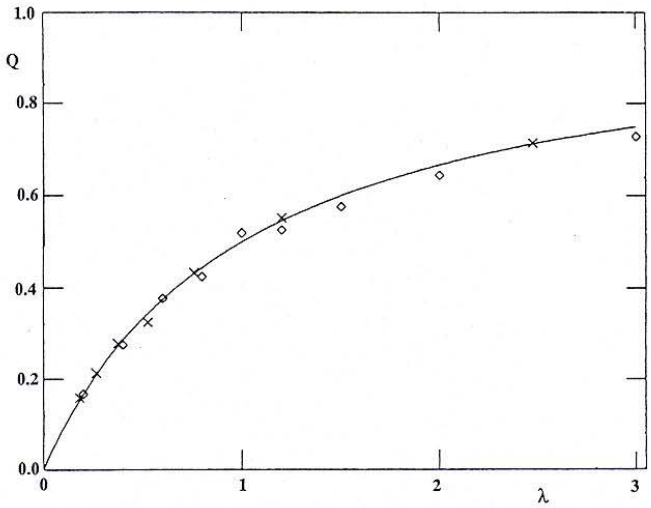

Figure 3. Average of $Q \simeq \langle q \rangle$ *vs* λ for $M = 200$. Diamonds: our model, with $N = 20$. Crosses: ZSP model, with $N = 10$, $\mu = 0.05$ and varying death probability δ. Averages over 10 000 generations.

The overlap $q^{\alpha\beta}$ represents the result of a comparison between two individuals. One may also envisage comparing three or more individuals together at a time. This defines a class of generalized overlaps which may also be explicitly calculated in our model.

6. Fluctuations and Ultrametricity. We have already seen, in Sections 3 and 4, that fluctuations appear in this model even in the limit $M \to \infty$. However, the quantity $Y(\tau)$, whose fluctuations have been computed in equations (20), (21) and (32), is not accessible if only the state of the population at a given generation is known. On the other hand, the average overlap $\langle q \rangle$ fluctuates during the evolution of the system, as shown by the simulation results of Fig. 6. In order to highlight this feature, let us compute, in the infinite genome limit $N \to \infty$, the root mean square deviation of Q. One has:

$$\overline{Q^2} = \overline{\left(\frac{1}{N} \sum_{i=1}^{N} \langle S_i \rangle^2 \right)^2} \simeq \overline{\langle q \rangle^2}. \tag{49}$$

In the infinite genome limit, we have seen that, if the relatedness of two individuals is equal to τ, their overlap q is given by:

$$q = e^{-\tau/\lambda} = e^{-4\mu t}. \tag{50}$$

On the other hand, the joint probability density $G(\tau, \tau')$ that two individuals,

α and β, have relatedness equal to τ, and two other ones, γ and δ, equal to τ', is given by:

$$G(\tau, \tau') = \left[\binom{M}{2}\right]^{-2} \overline{\sum_{(\alpha,\beta)} \vartheta_{\alpha\beta}(\tau) \sum_{(\gamma,\delta)} \vartheta_{\gamma\delta}(\tau')}, \tag{51}$$

where $\vartheta_{\alpha\beta}(\tau)$ has been defined in equation (18).

One obtains therefore:

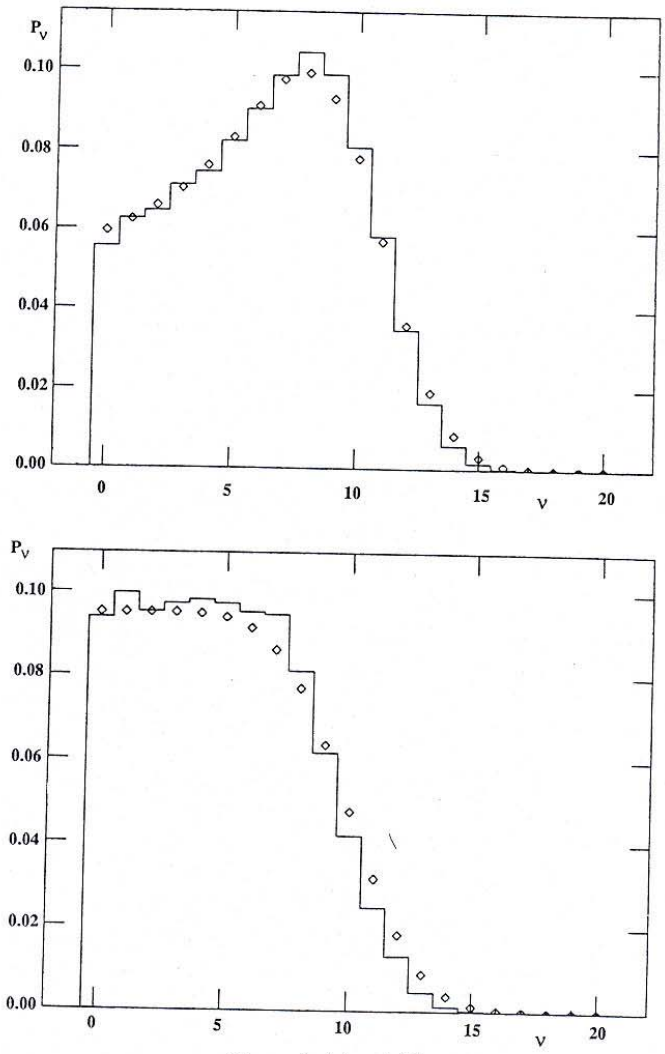

Figure 4. (a) and (b).

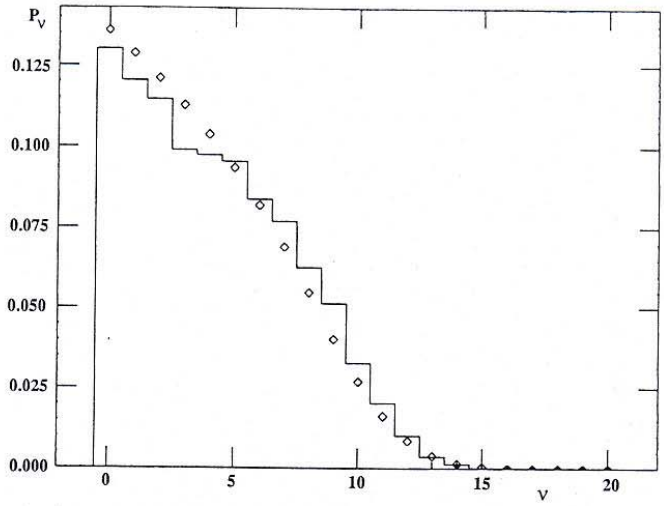

Figure 4. Histograms of the Hamming distance ν for $N=20$, $M=200$: (a) $\lambda=0.6$; (b) $\lambda=1.0$; (c) $\lambda=1.5$. Diamonds: equation (42). Averages over 10 000 generations.

$$\overline{\langle q\rangle^2}=\int_0^1 dq\int_0^1 dq'\int_0^\infty d\tau\int_0^\infty d\tau' G(\tau,\tau')qq'\,\delta(q-e^{-\tau/\lambda})\,\delta(q'-e^{-\tau'/\lambda}). \quad (52)$$

We now evaluate $G(\tau,\tau')$. One may easily convince oneself that:

$$G(\tau,\tau')=\theta(\tau-\tau')[2A(\tau,\tau')+4B(\tau,\tau')+C(\tau,\tau')] + (\tau\leftrightarrow\tau')+\delta(\tau-\tau')4D(\tau), \quad (53)$$

where A, B, C, D, are the probabilities of the genealogies shown in Figs 7a–d respectively, and the numerical factors take multiplicities into account. A simple calculation yields:

$$2A(\tau,\tau')=e^{-\tau-5\tau'}-e^{-3\tau-3\tau'}, \quad (54)$$

$$4B(\tau,\tau')=\tfrac{4}{3}(e^{\tau-2\tau'}-e^{-\tau-5\tau'}), \quad (55)$$

$$C(\tau,\tau')=e^{-3\tau-3\tau'}, \quad (56)$$

$$4D(\tau)=\tfrac{2}{5}e^{-\tau}-\tfrac{2}{3}e^{-3\tau}+\tfrac{4}{15}e^{-6\tau}. \quad (57)$$

We thus obtain:

$$\overline{\langle q\rangle^2}=\int_0^\infty d\tau\int_0^\infty d\tau' e^{-(\tau+\tau')/\lambda}G(\tau,\tau') =\frac{2\lambda^2(9\lambda^2+18\lambda+4)}{(\lambda+1)(\lambda+2)(3\lambda+2)(6\lambda+2)}. \quad (58)$$

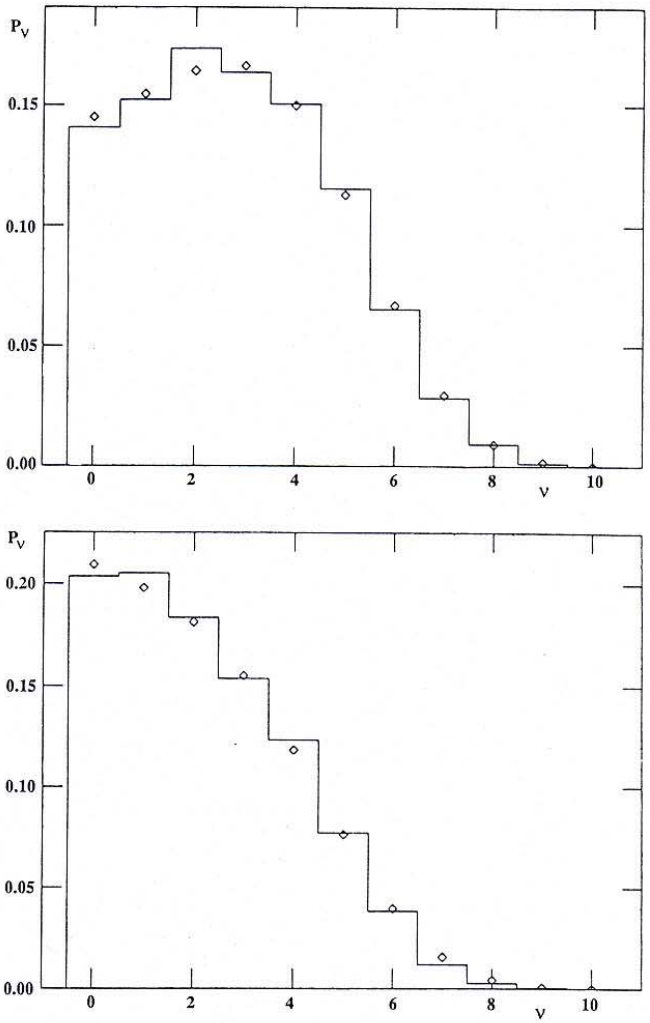

Figure 5. Histograms of the Hamming distance ν for the ZSP model, with $N=10$, $M=200$, and $\mu=0.05$. (a) $\delta=0.7$; (b) $\delta=0.8$. Diamonds: equation (42). Averages over 10 000 generations.

This implies:

$$\Delta Q^2=\overline{Q^2}-\bar{Q}^2\simeq\overline{\langle q\rangle^2}-\overline{\langle q\rangle}^2$$
$$=\frac{2\lambda^3}{(\lambda+1)^2(\lambda+2)(3\lambda+1)(3\lambda+2)}. \tag{59}$$

These results may be compared to those obtained by Stewart (1976) in an

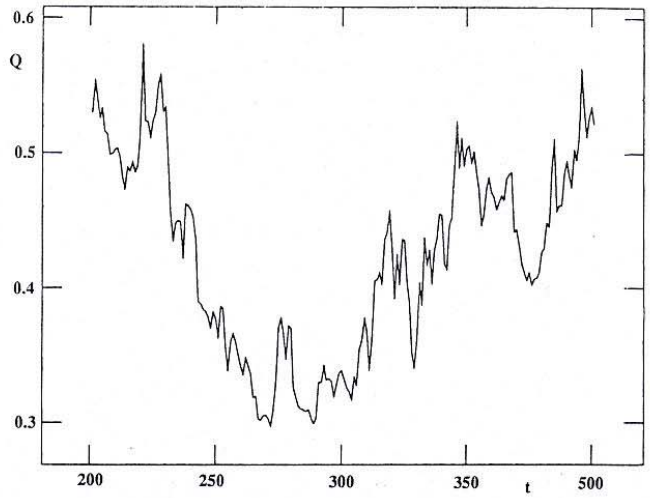

Figure 6. $Q(t)$ *vs* the generation number t for a single run of simulation, with $N=20$, $M=200$ and $\lambda=1$.

analogous problem. They hold in the infinite genome size limit. For comparison with the simulation data, it is important to consider the finite size corrections to ΔQ. A simple calculation shows that, to first order in $1/N$, ΔQ^2 should be increased by the following term:

$$\Delta Q'^2 = \frac{2\lambda^2(11\lambda+12)(2\lambda+1)}{3(\lambda+1)(\lambda+2)(3\lambda+1)(3\lambda+2)}\frac{1}{N}. \tag{60}$$

Figure 8 shows a plot of ΔQ *vs* λ for our model. The full line corresponds to the limit $N\to\infty$, whereas the dashed line takes into account the finite size correction equation (60).

The relatedness $\tau_{\alpha\beta}$ provides a natural notion of distance in populations of reproducing entities. This distance is obviously *ultrametric*, i.e. hierarchical, as witnessed by the very existence of genealogical trees. This corresponds to the fact that the usual triangle property of a distance:

$$\tau_{\alpha\beta} \leqslant \tau_{\alpha\gamma} + \tau_{\gamma\beta}, \qquad \forall\alpha, \beta, \gamma \in \Omega, \tag{61}$$

is replaced by the stronger property:

$$\tau_{\alpha\beta} \leqslant \max(\tau_{\alpha\gamma}, \tau_{\gamma\beta}), \qquad \forall\alpha, \beta, \gamma \in \Omega. \tag{62}$$

It is easy to calculate, in our model, the joint probability density $P(\tau_{\alpha\beta}, \tau_{\beta\gamma}, \tau_{\gamma\alpha})$ function of the relatednesses of three individuals. By use of equation (16) we obtain in fact:

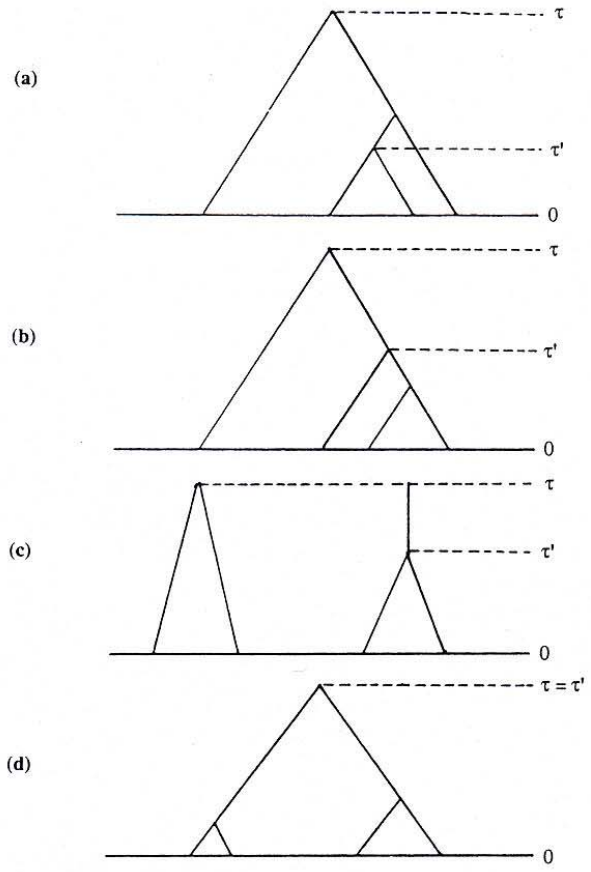

Figure 7. Genealogies involved in the evaluation of ΔQ. Integration on the unmarked time labels is understood.

$$P(\tau_{\alpha\beta}, \tau_{\beta\gamma}, \tau_{\gamma\alpha}) = \theta(\tau_{\alpha\beta} - \tau_{\beta\gamma})\, \delta(\tau_{\alpha\beta} - \tau_{\gamma\alpha}) e^{-\tau_{\alpha\beta} - 2\tau_{\beta\gamma}} + \text{Perm}(\alpha, \beta, \gamma). \quad (63)$$

Equation (50) implies that relatedness is unequivocally converted in genetic similarity in the infinite genome limit. In particular, if two individuals belong to the same τ-family, their overlap is not smaller than $e^{-\tau/\lambda}$. We may then derive from equation (39) the joint probability density $P(q^{\alpha\beta}, q^{\beta\gamma}, q^{\gamma\alpha})$ of the overlaps q, in the infinite genome limit $N \to \infty$:

$$P(q^{\alpha\beta}, q^{\beta\gamma}, q^{\gamma\alpha}) = \lambda^2 \theta(q^{\beta\gamma} - q^{\alpha\beta})\, \delta(q^{\alpha\beta} - q^{\gamma\alpha})\, (q^{\alpha\beta})^{\lambda-1} (q^{\gamma\alpha})^{2\lambda-1} + \text{Perm}(\alpha, \beta, \gamma). \quad (64)$$

By the same token, the distribution of τ-families $f(W_1, W_2, \ldots, W_k)$, given by equation (30), is converted into a cluster distribution in genome space. Since τ

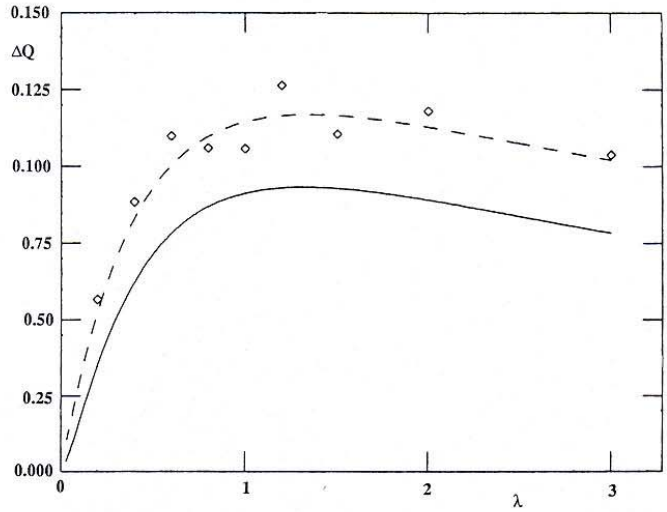

Figure 8. Root mean square deviation ΔQ of Q *vs* λ for $N=20$ and $M=200$. Averages over 10 000 generations. The full line is the theory equation (59), valid in the limit $N\to\infty$, while the dashed line takes into account the $1/N$ corrections equation (60).

and q are related by equation (50), is converted into a cluster distribution in genome space. Since τ and q are related by equation (50), these distributions describe the formation of clusters of genetically similar individuals.

When N is not strictly infinite, however, the mapping from genealogical relatedness to genetic similarity is no more deterministic. In particular, we do not expect the ultrametric structure of q appearing in equation (64) to remain strictly valid for finite N.

7. Genetic Drift. The time evolution of the average genotype $\langle S_i(t)\rangle$ is conveniently expressed by the correlation function $\chi(t)$, defined by:

$$\chi(t)=\frac{1}{N}\sum_{i=1}^{N}\overline{\langle S_i(t)\rangle\langle S_i(0)\rangle}. \tag{65}$$

For reasons that will appear evident in a short time, we shall indicate time dependences, in this section, by the generation number t instead of the reduced one τ.

We expect in general an exponential decay of correlations:

$$\chi(t)\propto e^{-2\mu^* t}, \tag{66}$$

which defines the effective mutation rate μ^*.

This correlation function is easily calculated. Let us consider an individual,

α, existing at generation 0, and an individual, β, existing at a following generation t. Let us denote by β' the ancestor of β, existing at generation 0. The relatedness between α and β' is distributed according to the probability density $p(\tau)$ obtained in equation (15). Given a relatedness τ between α and β', the individuals α and β are connected by a random walk in genome space, lasting $2\tau M+t$ generations. Their expected overlap is therefore equal to $e^{-2\mu(2\tau M+t)}$. We therefore have:

$$\chi(t)=\int_0^\infty d\tau e^{-\tau} e^{-2\mu(2\tau M+t)}=\frac{\lambda}{\lambda+1} e^{-2\mu t}. \tag{67}$$

Thus the effective mutation rate μ^*, equal to half the inverse characteristic time of correlation decay, is equal to the "bare" mutation rate μ. This is a well known property of neutral evolution (e.g. p. 47 of Kimura, 1983). We remark, in particular, that μ^* is independent of the population size M, when μ is fixed. We report in Fig. 9 a plot of the effective mutation rate μ^* *vs* the bare one μ for $N=20$ and different population sizes M.

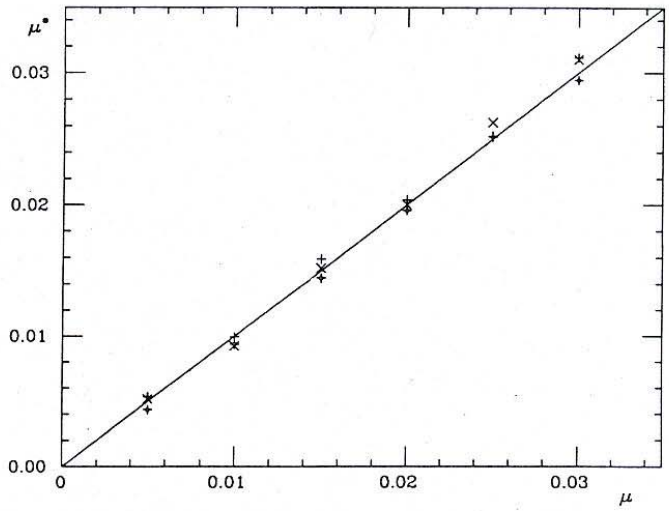

Figure 9. Effective mutation rate μ^* *vs* μ for $N=20$ and different population sizes M. Averages over 40 000 generations. Pluses: $M=10$; crosses: $M=80$; pluses with a hole: $M=160$. Line: $\mu^*=\mu$.

8. Discussion. The most obvious limitation of our model lies in the consideration of a flat fitness landscape. Our results, however, may have interesting implications for models of evolution in rugged fitness landscapes (Kauffman and Levin, 1987; Kauffman, 1989; Macken and Perelson, 1989). In

these models the fitness $H(\mathbf{S})$ of a given genome configuration $\mathbf{S}$ is picked at random from an identical probability distribution, independently for each of the 2^N ones. In the theory of disordered systems, such a model is known as the Random Energy Model (Derrida, 1980; 1981).

The behavior of the model is simplest in the case of sharp selection (corresponding to the $\beta \to \infty$ limit in the language of Amitrano *et al.*, 1989). Genotypes $\mathbf{S}$, whose fitness $H(\mathbf{S})$ does not exceed a threshold H_0 are unviable, and the corresponding individuals are removed outright from the population. Such genotypes are called forbidden: with our hypotheses, they are distributed at random on the hypercube of all possible genome configurations. We shall denote by x the fraction of forbidden genome configurations.

Let us consider the following form of the infinite genome size limit (p. 236 of Kimura, 1983):

$$N \to \infty, \qquad \sigma = \mu N = \text{const.} \tag{68}$$

Each mutation produces a new genome configuration, which is a forbidden one with probability x. In this case, the mutated individual dies, and one of the other ones reproduces to fill in the gap in the population. Thus the model becomes a form of the ZSP model, with a death probability δ given by:

$$\delta = \sigma x. \tag{69}$$

We expect that the genealogy statistics will be the same as in the neutral model, up to a suitable choice of the rescaling parameter κ.

However, we expect deviations in the statistics of the overlap q as well as in the genetic drift. It is easy to convince oneself that the decay of the autocorrelation $\chi(t)$ of the average genome will be still characterized by the same time constant as that of the drift of a single lineage in genome space. But this drift is constrained to take place only on allowed configurations, whose fitness $H(\mathbf{S})$ satisfies $H(\mathbf{S}) \geqslant H_0$. The problem reduces to that of a random walk on a hypercube with randomly distributed holes. This problem has recently been attentively investigated (e.g. Flesselles, 1989; Flesselles and Botet, 1989): it appears that, if the fraction x of holes is small, the walkers diffuse with a reduced, x-dependent, diffusion constant. But, at higher values of x, the existence of a stretched-exponential diffusion regime has been suggested. When x increases beyond a percolation threshold, one would go to a "trapped" regime, where memory of the initial configuration never disappears (Amitrano *et al.*, 1988). An explicit solution of this diffusion problem would immediately yield predictions for $\bar{Q}$, P_v etc., in the same way as $\phi_v(t)$ allowed us to draw predictions in the flat landscape case.

In spite of its simplicity, we believe that our model captures a few features of general relevance in evolving populations. In particular, the consideration of

quantities like $Y(\tau)$, the number of τ-families $F(\tau)$, or the family distribution function $f(W_1, W_2, \ldots, W_k)$, may be useful to describe the genetic structure of the population via the mapping of relatedness to similarity provided by equation (50). They could remain meaningful even in more sophisticated models of evolution. In the same way, the lack of the self-averaging property of some population averages may have a more general validity.

It is interesting to remark that our model exhibits most of the intriguing features of mean field spin glasses, such as lack of self-averaging, ultrametricity, nontrivial overlap distribution etc. It may therefore be useful as a pedagogical introduction to the new ideas in the theory of disordered systems.

On the other hand, the model predicts extremely large variabilities for any reasonably sized natural population—and in particular for the populations of *in vitro* replicating polynucleotides investigated by Biebricher and collaborators (e.g. the short review of Biebricher, 1986). Two effects conspire to reduce the variability of real populations: on the one hand, Natural Selection, whose neglect is the major weakness of the present model. The study of the behavior of the model in nontrivial fitness landscapes appears therefore as a most necessary step. On the other hand, populations like bacteria on a Petri dish, or *in vitro* replicating polynucleotides, are studied only a few generations after their foundation by one or few individuals. They are therefore far from equilibrium, as defined in the present model, which takes place only after a number of generations of order M. One is therefore led to consider the effects of a rapidly increasing population size. We expect novel features to appear when either of the effects just mentioned is introduced. For example, Weinberger (1987) has shown that a dynamical phase transition may appear in evolution models, as a function of the rate of population increase; whereas Amitrano *et al.* (1989) have discussed the effects of landscape correlation on the level of adaptation (see also Kauffman, 1989).

We hope that the present model could be considered as a useful stepping stone towards the study of these more complicated situations.

B. D. thanks T. Spencer for his kind hospitality at the Institute for Advanced Studies. L. P. acknowledges the support of the Centre National pour la Recherche Scientifique. He thanks J. Prost and the Groupe de Physico-Chimie Théorique de l'E.S.P.C.I. for a most pleasant hospitality. Both authors thank M. Mézard for illuminating discussions. L. P. also thanks G. Ciccotti for useful suggestions, and S. Nicolis for help in the simulations.

APPENDIX 1

We report here, for completeness, the calculation of $X_{n_1 n_2 \ldots n_k}(\tau)$. We have seen that the probability of a given genealogy in $\Pi_n(\tau_n, \tau_{n-1}, \ldots, \tau_k)$ and depends only on n, the number of

individuals, k, the number of families, and on the branching epochs $\tau_n, \tau_{n-1}, \ldots, \tau_k$. We define a *genealogical scheme* as a set of genealogical trees in which the order (but not the epochs) of branching of the different lineages is fixed. The probability $P^*(\tau)$ of a given genealogical scheme depends only on the number n of individuals involved, on the number k of their ancestors, and on the time span τ, and is given by:

$$P^*(\tau)=\int_0^\tau d\tau_n \int_{\tau_n}^\tau d\tau_{n-1} \cdots \int_{\tau_{k+1}}^\tau d\tau_k \prod_{l=k}^{n} \exp\left[-\frac{l(l-1)}{2}(\tau_l-\tau_{l+1})\right]. \tag{A1.1}$$

We have set $\tau_{n+1}=0$. This probability has the form of a Laplace convolution, and its Laplace transform is readily calculated:

$$\tilde{P}^*(\beta)=\prod_{l=k}^{n}\left[\beta+\frac{l(l-1)}{2}\right]^{-1}, \tag{A1.2}$$

where we denote by β the Laplace parameter.

The quantity we are looking for is the product of this probability for the number of genealogical schemes of n individuals divided in k families, such that the first n_1 ones belong to the first one, the second n_2 ones to the second one, . . . and the last n_k ones to the k^{th} one. Let us denote this number by $G_{n_1 n_2 \ldots n_k}$. For a single family, G_n is readily calculated by induction. Consider a genealogical tree of n individuals, and imagine cutting it at the level of the most recent common ancestor of two individuals. We obtain a tree for $n-1$ individuals. Since the two individuals involved may be chosen in $n(n-1)/2$ ways, we obtain:

$$G_n=\frac{n!(n-1)!}{2^{n-1}}. \tag{A1.3}$$

Let us now consider a set of k genealogical trees. In order to get a genealogical scheme, we must specify the order in which the different branching times are arranged among the different trees. The order within a single genealogical tree is fixed by definition. The number of possible arrangements of the $n-k$ branching times is equal to the number of combinations of the $n-k$ order labels, such that the first tree has n_1-1 of them, the second one n_2-1, . . . and the last one n_k-1 of them. We have therefore:

$$\begin{aligned} G_{n_1 n_2 \ldots n_k} &= \frac{(n-k)!}{(n_1-1)!(n_2-1)!\ldots(n_k-1)!}\prod_{l=1}^{k} G_{n_l} \\ &= \frac{(n-k)!}{2^{n-k}} n_1!n_2!\ldots n_k!. \end{aligned} \tag{A1.4}$$

We can now write down the Laplace transform of $X_{n_1 n_2 \ldots n_k}(\tau)$:

$$\tilde{X}_{n_1 n_2 \ldots n_k}(\beta)=n_1!n_2!\ldots n_k!\,\frac{(n-k)!}{2^{n-k}}\prod_{l=k}^{n}\left[\beta+\frac{l(l-1)}{2}\right]^{-1}. \tag{A1.5}$$

By inverting the Laplace transformation we recover equation (17).

APPENDIX 2

We derive here, following Derrida and Bessis (1988), the expression equation (26) of the probability $Z_k(\tau)$ that the whole population is divided in k τ-families.

We first consider the probability $Q_k^{(n)}(\tau)$ that n individuals belong to k τ-families. One has of course:

$$Z_k(\tau)=\lim_{n\to\infty} Q_k^{(n)}(\tau). \tag{A2.1}$$

The time evolution of the $Q_k^{(n)}$ is given by:

$$\frac{dQ_k^{(n)}}{d\tau}=-\frac{k(k-1)}{2}Q_k^{(n)}+\frac{k(k+1)}{2}Q_{k+1}^{(n)}, \tag{A2.2}$$

with the initial condition:

$$Q_k^{(n)}(0)=\delta_{kn}. \tag{A2.3}$$

Equation (A2.2) may be solved recursively, by going to the Laplace transform, and taking into account the initial condition (A2.3). The result reads:

$$Q_k^{(n)}(\tau)=\frac{n!(n-1)!}{k!(k-1)!}\sum_{l=k}^{n}(-1)^{k+l}\frac{(2l-1)(k+l-2)!}{(l-k)!(n+l-1)!(n-l)!}\exp\left[-\frac{l(l-1)}{2}\tau\right], \tag{A2.4}$$

and equation (27) follows by taking the limit $n\to\infty$.

The moments of $F(\tau)$ can be computed straightforwardly. We have, for example:

$$\Phi(\tau)=\sum_{k=1}^{\infty}kZ_k(\tau)=\sum_{k=1}^{\infty}\sum_{p=k}^{\infty}(-1)^{p+k}(2p-1)\frac{(p+k-2)!}{[(k-1)!]^2(p-k)!}\exp\left[-\frac{p(p-1)}{2}\tau\right]. \tag{A2.5}$$

By rearranging the sum we obtain:

$$\Phi(\tau)=\sum_{p=1}^{\infty}(-1)^p(2p-1)A_p\exp\left[-\frac{p(p-1)}{2}\tau\right], \tag{A2.6}$$

where:

$$A_p=\sum_{k=1}^{p}(-1)^k\frac{(p+k-2)!}{[(k-1)!]^2(p-k)!}=(-1)^p, \tag{A2.7}$$

and equation (27) follows directly. In the same way we obtain:

$$\overline{F^2(\tau)}=\sum_{p=1}^{\infty}(2p-1)(p^2-p+1)\exp\left[-\frac{p(p-1)}{2}\tau\right]. \tag{A2.8}$$

LITERATURE

Amitrano, C., L. Peliti and M. Saber. 1988. Neutralism and adaptation in a simple model of molecular evolution. *C. R. Acad. Sci. Paris, Sér. III* **307**, 803–806.

Amitrano, C., L. Peliti and M. Saber. 1990. Population dynamics in a spin-glass model of chemical evolution. *J. molec. Evol.*, **29**, 513–525.

Biebricher, C. K. 1986. Darwinian evolution of self-replicating RNA. *Chemica Scripta* **26B**, 51–57.

Demetrius, L., P. Schuster and K. Sigmund. 1985. Polynucleotide evolution and branching processes. *Bull. math. Biol.* **47**, 239–262.

Derrida, B. 1980. Random-energy model—limit of a family of disordered systems. *Phys. Rev. Lett.* **45**, 79–82.

Derrida, B. 1981. Random-energy model—an exactly solvable model of disordered systems. *Phys. Rev.* **B24**, 2613–2626.

Derrida, B. and D. Bessis. 1988. Statistical properties of valleys in the annealed random map model. *J. Phys.* **A21**, L509–L515.

Derrida, B. and H. Flyvbjerg. 1987. Statistical properties of randomly broken objects and of multivalley structures in disordered systems. *J. Phys.* **A20**, 5273–5288.

Ebeling, W. and R. Feistel. 1977. Stochastic theory of a molecular replication process with selection character. *Ann. Phys. (Leipzig)* **34**, 81–90.

Eigen, M., J. McCaskill and P. Schuster. 1988. Molecular quasi-species. *J. phys. Chem.* **92**, 6881–6891.

Flesselles, J.-M. 1989. Contribution théorique à l'étude de la relaxation dans les verres de spin. Thèse, Université Paris-Sud.

Flesselles, J.-M. and R. Botet. 1989. Derivation of a stretched-exponential time relaxation. *J. Phys.* **A22**, 903–909.

Fontana, W., W. Schnabl and P. Schuster. 1989. Physical aspects of evolutionary optimization and adaptation. *Phys. Rev.* **A40**, 3301–3321.

Fontana, W. and P. Schuster. 1987. A computer model of evolutionary optimization. *Biophys. Chem.* **26**, 123–147.

Hofbauer, J. and K. Sigmund. 1988. *The Theory of Evolution and Dynamical Systems—Mathematical Aspects of Selection*. Cambridge: Cambridge University Press.

Kauffman, S. A. 1989. Adaptation on rugged fitness landscapes. In *Complex Systems. SFI Studies in the Science of Complexity*, D. Stein (ed.), pp. 527–617. Reading, MA: Addison-Wesley.

Kauffman, S. A. and S. Levin. 1987. Towards a general theory of adaptive walks in rugged fitness landscapes. *J. theor. Biol.* **128**, 11–45.

Kimura, M. 1983. *The Neutral Theory of Molecular Evolution*. Cambridge: Cambridge University Press.

Kimura, M. and J. F. Crow. 1964. The number of alleles that can be maintained in a finite population. *Genetics* **49**, 725–738.

Macken, C. A. and A. S. Perelson. 1989. Protein evolution in rugged fitness landscapes. *Proc. natl Acad. Sci. U.S.A.* **86**, 6191–6195.

Mézard, M., G. Parisi and M. A. Virasoro. 1987. *Spin-Glass Theory and Beyond*. Singapore: World Scientific.

Nowak, M. and P. Schuster. 1989. Error thresholds of replication in finite populations. *J. theor. Biol.* **137**, 375–395.

Schuster, P. and K. Sigmund. 1985. Dynamics of evolutionary optimization. *Ber. Bunsenges. phys. Chem.* **89**, 668–682.

Schuster, P. and J. Swetina. 1988. Stationary mutant distribution and evolutionary optimization. *Bull. math. Biol.* **50**, 635–660.

Stewart, F. M. 1976. Variability in the amount of heterozygosity, maintained by neutral mutations. *Theor. Pop. Biol.* **9**, 188–201.

Swetina, J. and P. Schuster. 1982. Self-replication with errors—a model for polynucleotide replication. *Biophys. Chem.* **16**, 329–345.

Weinberger, E. 1987. A model of natural selection that exhibits a dynamic phase transition. *J. stat. Phys.* **49**, 1011–1028.

Wright, S. 1937. The distribution of gene frequencies in populations. *Proc. natl Acad. Sci. U.S.A.* **23**, 307–320.

Zhang, Y.-C., M. Serva and M. Polikarpov. 1990. Diffusion-reproduction processes. *J. stat. Phys.*, in press.

Received 10 January 1990
Revised 2 April 1990