Stochastic model of evolving populations

1998 J. Phys. A: Math. Gen. 31 417

(http://iopscience.iop.org/0305-4470/31/2/004)

J. Phys. A: Math. Gen. **31** (1998) 417–429. Printed in the UK PII: S0305-4470(98)84712-4

Stochastic model of evolving populations

D Bonnaz and A J Koch

Institut de physique expérimentale, Université de Lausanne, CH-1015 Lausanne, Switzerland

Received 4 June 1997, in final form 20 August 1997

Abstract. A general formalism is described which allows the study of a wide range of stochastic problems related to evolution. We apply it to the treatment of the evolution of a finite population in the presence of mutations and selective advantage. The results are close to those of Eigen and Schuster's error threshold model, in the presence of randomly drawn sterile sequences. The theoretical results are in excellent agreement with numerical simulations.

1. Introduction

Eigen's (1971) deterministic model for the prebiotic evolution of macromolecules has had a considerable impact on the way we consider evolutionary problems from a theoretical standpoint. It is an outstanding example of what a model can bring to the understanding of biological phenomena. For instance, it shone light onto the existence of an *error threshold*: in the presence of mutations, information is conserved in a population of macromolecules only if the mutation rate μ is smaller than a critical value μ_c. Beyond the error threshold, Eigen's model predicts that the concentrations of all possible types of macromolecules become equal. This result is, however, unrealistic for the following reason. Consider macromolecules made of 200 monomers which can be of four types (A, T, G and C). The number of distinct molecules is then equal to $4^{200} \approx 10^{120}$; this number largely exceeds the number of baryons in the known universe ($\approx 10^{80}$). This shows that, for biologically relevant parameters, the number of distinct macromolecules which could, in principle, be synthesized is tremendously larger than the number of macromolecules which can effectively be produced. As a consequence, the overwhelming majority of possible macromolecules are not represented in the population and never will be.

Although Eigen and co-workers (Eigen and Schuster, 1977, Eigen *et al* 1989, Schuster and Svetina 1988) are conscious of this problem, they have—to our knowledge—never proposed a modified version of their theory in order to take into account the finiteness of biological populations, even if Nowak and Schuster (1989) have proposed numerical simulations of Eigen's system in small populations.

Modelling evolving populations is of current interest; several recent papers deal with this topic (see Woodcock and Higgs 1996 or Prügel–Bennett 1997 and references therein). The model presented hereafter was originally developed in order to describe the evolution of a population composed of a finite number of individuals. Due to its probabilistic formulation, it is well suited to study stochastic effects in finite populations. Furthermore, we introduce *ab initio* the fact that a given fraction of individuals of the population are unable to reproduce. If, for instance, the model is applied to self-reproducing macromolecules, this takes into account the well known experimental fact that most of the RNA molecules cannot self-replicate (Joyce and Orgel 1993, Eigen 1993).

0305-4470/98/020417+13$19.50

The formalism presented here is adapted to the description of biological populations submitted to mutations and selection. From a theoretical point of view, this method lies midway between Eigen's deterministic model (which does not take care of the finiteness of the population size) and Kimura's (1962) theory of genetic drift which originally does not take mutations into account and only considers a small number of genotypes. The method is essentially based on the idea that the mean value of any quantity linked to the population's genomic distribution can be evaluated by calculating its average change between two successive generations of an arbitrary population and by weighting the result over the probability of occurrence of the former generation.

This paper is organized as follows. We shall first develop the formalism of the method. Once the theoretical framework is clarified, we shall study a model with selective advantage and lethal or sterile genotypes. The distribution of sterile genotypes introduces a kind of disorder in the sequence space. We shall discuss both cases of low disorder (few sterile genotypes) and strong disorder (few fertile genotypes).

2. Theoretical basis of the stochastic model

Consider a population $\mathcal{P}$ composed of N individuals which are arbitrarily numbered from 1 to N to distinguish them. The population is supposed to beget N children which will form the population $\mathcal{P}'$ of the next generation.

Each individual is characterized by its genome. To fix the ideas, we shall admit that a genome is a binary sequence composed of ν monomers. It can be represented by a vector $\boldsymbol{x} = \{x^1, x^2, \ldots, x^\alpha, \ldots, x^\nu\}$ with $x^\alpha \in \{-1, +1\}$ $(1 \leqslant \alpha \leqslant \nu)$ belonging to the sequence space $\mathcal{X}$. This set $\mathcal{X}$ of all distinct sequences $\boldsymbol{x}$ contains 2^ν elements.

We shall note $\mathcal{G}$ the set of genomes of all individuals belonging to $\mathcal{P}$: $\mathcal{G} = \{\boldsymbol{x}_i | 1 \leqslant i \leqslant N\} \in \mathcal{X}^N$ where $\boldsymbol{x}_i$ is the genome of individual i of the population. $\mathcal{G}$ is the gene pool of $\mathcal{P}$ (Maynard Smith 1983). Hereafter quantities with a prime (such as $\boldsymbol{x}'_i$) are related to the offspring population $\mathcal{P}'$, while quantities without a prime (as $\boldsymbol{x}_i$) concern the parental population $\mathcal{P}$.

Let us introduce the probability $W(\mathcal{G}', \mathcal{G})$ to get the gene pool $\mathcal{G}' = \{\boldsymbol{x}'_i | 1 \leqslant i \leqslant N\}$, if one generation earlier the pool was $\mathcal{G}$. We have obviously

$$\sum_{\mathcal{G}' \in \mathcal{X}^N} W(\mathcal{G}', \mathcal{G}) = 1$$

where the sum is taken over all possible pools $\mathcal{G}'$. $W(\mathcal{G}', \mathcal{G})$ describes the reproduction of the population; it takes care of effects such as selective advantages of some genotypes compared with others, or of mutations occurring during the replication of the genomes. We restrict ourselves to time-independent $W(\mathcal{G}', \mathcal{G})$.

By using $W(\mathcal{G}', \mathcal{G})$, it is possible to calculate the probability $P(\mathcal{G}', t+1)$ of having the pool $\mathcal{G}'$ at generation $t+1$ if the probability $P(\mathcal{G}, t)$ of observing $\mathcal{G}$ at generation t is known:

$$P(\mathcal{G}', t+1) = \sum_{\mathcal{G} \in \mathcal{X}^N} W(\mathcal{G}', \mathcal{G}) P(\mathcal{G}, t). \tag{1}$$

Let $F(\mathcal{G})$ be an arbitrary function of $\mathcal{G}$. We define the following two quantities:

$$\overline{F}(t) = \sum_{\mathcal{G} \in \mathcal{X}^N} P(\mathcal{G}, t) F(\mathcal{G}) \tag{2}$$

$$\langle F \rangle(\mathcal{G}) = \sum_{\mathcal{G}' \in \mathcal{X}^N} W(\mathcal{G}', \mathcal{G}) F(\mathcal{G}'). \tag{3}$$

$\overline{F}(t)$ corresponds to the *average value* of F on a population at generation t. Similarly, $\langle F\rangle(\mathcal{G})$ is the *expected value* of F estimated on a population if, one generation earlier, the gene pool was $\mathcal{G}$. By using (1)–(3), we see that

$$\overline{F}(t+1) = \overline{\langle F\rangle}(t). \tag{4}$$

If the system reaches a stationary state (where $P(\mathcal{G}, t+1) = P(\mathcal{G}, t)$), then the preceding relation leads to:

$$\overline{F}(t) = \overline{\langle F\rangle}(t). \tag{5}$$

Given $W(\mathcal{G}', \mathcal{G})$, this relation enables us to evaluate the average value of any function $F(\mathcal{G})$.

Now let us precise the form of $W(\mathcal{G}', \mathcal{G})$. The evolution of the population is given by the link between two successive generations. We make the following hypotheses which are close to those used by Serva and Peliti (1991).

- At each generation, the population is formed by N individuals.
- Successive generations do not overlap: all parents die before their offspring reproduce.
- Let $\mathcal{G}$ be the parental pool and $\mathcal{G}'$ be the offspring pool. The offspring population is obtained from the parental one by N independent reproduction events. Consequently, $W(\mathcal{G}', \mathcal{G})$ can be decomposed as

$$W(\mathcal{G}', \mathcal{G}) = \prod_{j=1}^{N} w(\boldsymbol{x}'_j, \mathcal{G}) \tag{6}$$

where $w(\boldsymbol{x}'_j, \mathcal{G})$ is the probability that the individual j belonging to $\mathcal{P}'$ gets the genome $\boldsymbol{x}'_j$ if the parental gene pool was $\mathcal{G}$. Of course the $w(\boldsymbol{x}'_j, \mathcal{G})$ are normalized:

$$\sum_{\boldsymbol{x}'_j \in \mathcal{X}} w(\boldsymbol{x}'_j, \mathcal{G}) = 1. \tag{7}$$

In order to specify $w(\boldsymbol{x}', \mathcal{G})$ more precisely, we consider for instance an asexual reproduction mode of a haploid species†.

Let us suppose that a given fitness $S(\boldsymbol{x}) \geqslant 0$ is associated to each genotype $\boldsymbol{x} \in \mathcal{X}$: an individual with genotype $\boldsymbol{x}$ is expected to beget $S(\boldsymbol{x})$ children. So, if we call $v(\boldsymbol{x}', \boldsymbol{x})$ the probability to get the genome $\boldsymbol{x}'$ as a copy (with eventual mutations) of $\boldsymbol{x}$, we have

$$w(\boldsymbol{x}', \mathcal{G}) = \frac{1}{\sum_{j=1}^{N} S(\boldsymbol{x}_j)} \sum_{i=1}^{N} S(\boldsymbol{x}_i) v(\boldsymbol{x}', \boldsymbol{x}_i) \tag{8}$$

since the N possibilities of drawing a particular parent $\boldsymbol{x}_i$ ($1 \leqslant i \leqslant N$) are mutually exclusive events. The probabilities $v(\boldsymbol{x}', \boldsymbol{x})$ verify the normalization relation

$$\sum_{\boldsymbol{x}' \in \mathcal{X}} v(\boldsymbol{x}', \boldsymbol{x}) = 1.$$

The decomposition (6) of $W(\mathcal{G}', \mathcal{G})$ has a practical interest. Consider a function $F(\boldsymbol{x}_\lambda)$ depending on the genome $\boldsymbol{x}_\lambda$ of a single individual λ. The average value $\langle F\rangle(\mathcal{G})$ is given by

$$\langle F\rangle(\mathcal{G}) = \sum_{\boldsymbol{x}'_\lambda \in \mathcal{X}} w(\boldsymbol{x}'_\lambda, \mathcal{G}) F(\boldsymbol{x}'_\lambda). \tag{9}$$

† As noticed by Kimura (1962), under certain conditions such a gene pool model also describes sexually reproducing populations. Similarly, the very same model can also be applied to haploid as to polyploid populations.

More generally, if F depends on the genomes $\boldsymbol{x}_{\lambda_1}, \boldsymbol{x}_{\lambda_2} \ldots \boldsymbol{x}_{\lambda_k}$ with k distinct but arbitrarily chosen indices $\lambda_1,\ldots,\lambda_k$. The average value $\langle F\rangle(\mathcal{G})$ is then given by

$$\langle F\rangle(\mathcal{G}) = \sum_{\boldsymbol{x}'_{\lambda_1}\in\mathcal{X}} \sum_{\boldsymbol{x}'_{\lambda_2}\in\mathcal{X}} \cdots \sum_{\boldsymbol{x}'_{\lambda_k}\in\mathcal{X}} \left(\prod_{a=1}^{k} w(\boldsymbol{x}'_{\lambda_a}, \mathcal{G})\right) F(\boldsymbol{x}'_{\lambda_1}, \ldots, \boldsymbol{x}'_{\lambda_k}). \tag{10}$$

Equations (5) and (10) form the core of the method used here. It is possible to evaluate the time evolution of any function if the transition probabilities $w(\boldsymbol{x}'_\beta, \mathcal{G})$ are known.

3. Evolution in the presence of selective advantage

Let us now apply the previous formalism to the evolution of a population $\mathcal{P}$ of N haploid individuals reproducing asexually. We suppose that the different genomes can be sorted in two classes: the fertile and the sterile (or lethal) ones and we assume that both are uniformly distributed in the sequence space. To take this into account, we characterize once for all (quenched disorder) each genotype $\boldsymbol{x} \in \mathcal{X}$ by the value of a fertility function $f(\boldsymbol{x})$ in the following way:

$$f(\boldsymbol{x}) = \begin{cases} 1 \text{ (for fertile)} & \text{with probability } p \\ 0 \text{ (for sterile)} & \text{with probability } 1-p. \end{cases}$$

(See also Peliti and Bastolla (1994)). A second element is directly inspired from Eigen's models. We shall assume that the genotype $\mathbf{1} = (1,\ldots,1)$ has a selective advantage s compared with all other genotypes in $\mathcal{X}$. Naturally, $f(\mathbf{1}) = 1$ (it would be nonsense to give a selective advantage to a lethal genotype). Given the parental population $\mathcal{P}$, the possible values that the genome $\boldsymbol{x}'_j$ of an arbitrary individual belonging to the offspring population $\mathcal{P}'$ are enumerated below with their respective probabilities:

$$\boldsymbol{x}'_j = \begin{cases} \boldsymbol{x}_i & \frac{f(\boldsymbol{x}_i)}{n_f+sn_o}(1-\mu\nu) \\ M_\alpha \boldsymbol{x}_i & \frac{f(\boldsymbol{x}_i)}{n_f+sn_o}\mu \\ \mathbf{1} & \frac{sn_0}{n_f+sn_o}(1-\mu\nu) \\ M_\alpha \mathbf{1} & \frac{sn_0}{n_f+sn_o}\mu \end{cases} \tag{11}$$

with $1 \leqslant i \leqslant N$. The various terms appearing therein will be explained below. For each monomer x_i^α of $\boldsymbol{x}_i$, there is a small probability μ to be badly copied. We assume here that μ is small enough so the eventuality of multiple errors during the replication of $\boldsymbol{x}$ can be neglected. $M_\alpha \boldsymbol{x}_i$ corresponds to the genome obtained by making a single mutation in $\boldsymbol{x}_i$ at position $\alpha \in \{1,\ldots,\nu\}$ during its replication. The probabilities listed in (11) correspond to the term $S(\boldsymbol{x}_i)v(\boldsymbol{x}', \boldsymbol{x}_i)$ appearing in (8). The probability to draw a sterile individual is zero.

In expression (11), n_f corresponds to the number of fertile individuals in the parental population and n_o to the number of parents with genotype $\mathbf{1}$. By definition, n_f and n_o are respectively equal to

$$n_f = \sum_{i=1}^{N} f(\boldsymbol{x}_i) \qquad \text{and} \qquad n_o = \sum_{i=1}^{N} \delta(\boldsymbol{x}_i - \mathbf{1})$$

where

$$\delta(\boldsymbol{x} - \mathbf{1}) = \prod_{\alpha=1}^{\nu} \delta_1^{x^\alpha}.$$

$\delta_1^{x^\alpha}$ is the Kronecker delta. Equation (11) only makes sense if $n_f > 0$, which means that there is at least one fertile individual in the parental population; if this was not the case, the population would become extinct.

By assuming that the population has reached a dynamical equilibrium, one determines without difficulties the mean number $\overline{n_f}$ of fertile individuals and $\overline{n_o}$, the mean number of individuals bearing the genomic sequence **1**; $\overline{n_f}$ and $\overline{n_o}$ are solutions of

$$\overline{\langle n_f \rangle}(t) = \overline{n_f}(t) \qquad \text{and} \qquad \overline{\langle n_o \rangle}(t) = \overline{n_o}(t).$$

Neglecting back-mutations on **1** and fluctuations for the calculus of $\overline{n_o}$, the solutions of these equations are respectively

$$\overline{n_f} = N[1 - (1-p)\mu\nu] \tag{12}$$

$$\overline{n_o} = \begin{cases} N\left(1 - \dfrac{s+p}{s}\mu\nu\right) & \text{if } \mu < \mu_c = \frac{s}{\nu(s+p)} \\ 0 & \text{otherwise.} \end{cases} \tag{13}$$

The solution for $\overline{n_o}$ shows that, according to the mutation rate μ, the population dynamics presents essentially two distinct behaviours.

• If $\mu < \mu_c = s/(\nu(s+p))$, there are, on average, $\overline{n_o} > 0$ individuals belonging to the master type **1**. We shall see hereafter that the whole population, although submitted to genetic drift, remains localized in the sequence space in the neighbourhood of **1**.

• For $\mu \geqslant \mu_c$, the mutational load is so heavy that the genotype **1**, despite its selective advantage, disappears; if **1** appears by chance, it will vanish some generations later, so that $\overline{n_o} = 0$ if $\mu \geqslant \mu_c$.

The mean fitness f_m of the population is defined by:

$$f_m = \frac{1}{N}\sum_{i=1}^{N}[1 + s\delta(\boldsymbol{x}_i - \mathbf{1})]f(\boldsymbol{x}_i).$$

Using (12) and (13), one obtains for its average

$$\overline{f_m} = \begin{cases} (1+s)(1-\mu\nu) & \text{if } \mu < \mu_c = \frac{s}{\nu(s+p)} \\ [1-(1-p)\mu\nu] & \text{otherwise.} \end{cases}$$

$\overline{f_m}$ is a continuous and decreasing function of μ. If $\mu < \mu_c$, $\overline{f_m}$ does not depend on the fraction p of fertile genotypes: the decrease of mean fitness due to the presence of sterile individuals is exactly compensated by the increase of $\overline{f_m}$ due to the existence of fitter individuals with genotype **1**. This result, although surprising, is common, see Higgs (1994) for other examples. The presence of lethal genotypes has another important consequence: it increases the value of μ_c, allowing larger mutation rates in the population without affecting $\overline{f_m}$.

To study the change in the population dynamics around μ_c, we have calculated the average variance $\overline{\sigma^2}$ of the population and the mean square shift $\overline{\langle(\Delta \boldsymbol{X})^2\rangle}$ of the mean genotype $\boldsymbol{X} = (1/N)\sum_{i=1}^{N}\boldsymbol{x}_i$ between two successive generations. To determine these quantities, we shall need the average values of the following five terms: $\sum \boldsymbol{x}_i$, $\sum f(\boldsymbol{x}_i)\boldsymbol{x}_i$, $\sum \boldsymbol{x}_i\boldsymbol{x}_j$, $\sum f(\boldsymbol{x}_i)\boldsymbol{x}_i\boldsymbol{x}_j$ and $\sum f(\boldsymbol{x}_i)f(\boldsymbol{x}_j)\boldsymbol{x}_i\boldsymbol{x}_j$. Detailed calculations will only be given for the first two expressions; the other ones are found by using similar reasoning. Let us start with two remarks.

• As written above, the fertile genotypes are uniformly distributed with frequency p in the sequence space $\mathcal{X}$. So, a term like

$$\sum_{\alpha=1}^{\nu} f(M_\alpha \boldsymbol{x}_j) M_\alpha \boldsymbol{x}_j$$

which will often appear in the calculations can be approximated by its mean value on the phase space (as if we were in the presence of annealed disorder):

$$\sum_{\alpha=1}^{\nu} f(M_\alpha \boldsymbol{x}_j) M_\alpha \boldsymbol{x}_j \approx (\nu - 2) p \boldsymbol{x}_j.$$

This approximation is good as long as $\nu p \gg 1$ (ν is the number of first neighbours of an arbitrary sequence $\boldsymbol{x}$ in $\mathcal{X}$ and νp corresponds to the average number of fertile neighbours of $\boldsymbol{x}$). If $\nu p < 1$, this approximation becomes doubtful.

• We shall also estimate the fluctuations of n_f and n_o by evaluating the variances of these two quantities. We obtain that for n_f the fluctuations are of order $\sqrt{N}$. For n_o the calculations are more delicate. Supposing that the fluctuations are small compared with n_o, we obtain that they are also of order $\sqrt{N}$. Thus, with the assumption that N is sufficiently large, we can consider n_f and n_o as constants whose values are given by (12) and (13) respectively. Hereafter, we shall write n_f instead of $\overline{n_f}$ and n_o rather than $\overline{n_o}$ in order to simplify the notations. The most visible finite-size effect is the shift of the error threshold. Indeed a population with n_o of order $\sqrt{N}$ is likely to escape from the master genotype **1** in few generations. Consequently, we deduce from (13) that μ_c has to be replaced by $\mu_c(N)$ determined by

$$\mu_c(N) = \mu_c - \frac{C\mu_c}{\sqrt{N}}$$

where C is a constant. On the other hand, numerical simulations show that the lifetime of a master population with n_o of order N increases exponentially with N, so it is very improbable to observe any extinction during one experiment. This is why one is justified to talk of an error threshold despite the finiteness of the lifetime of the master sequence.

So, let us begin with the evaluation of the average value of $\sum f(\boldsymbol{x}_j)\boldsymbol{x}_j$. The transition probabilities are given by (11).

$$\begin{aligned}\left\langle \sum_{i=1}^{N} f(\boldsymbol{x}_i')\boldsymbol{x}_i' \right\rangle &= \sum_{i=1}^{N} \frac{1}{n_f + sn_o} \sum_{j=1}^{N} [(1-\mu\nu) f(\boldsymbol{x}_j)\boldsymbol{x}_j \\ &\quad + \mu \sum_{\alpha=1}^{\nu} f(M_\alpha \boldsymbol{x}_j) M_\alpha \boldsymbol{x}_j] \cdot (1 + s\delta(\boldsymbol{x}_j - \mathbf{1})) \cdot f(\boldsymbol{x}_j) \\ &= \frac{n_f - 2\mu N p}{n_f + sn_o} \left[sn_o \mathbf{1} + \sum_{j=1}^{N} f(\boldsymbol{x}_j)\boldsymbol{x}_j \right].\end{aligned}$$

By using (5), one obtains the average value of this quantity:

$$\sum_{i=1}^{N} \overline{f(\boldsymbol{x}_i)\boldsymbol{x}_i} = \frac{n_f - 2\mu N p}{sn_o + 2\mu N p} sn_o \mathbf{1}.$$

We have also for each i

$$\langle \boldsymbol{x}_i' \rangle = \frac{1 - 2\mu}{n_f + sn_o} \left[sn_o \mathbf{1} + \sum_{j=1}^{N} f(\boldsymbol{x}_j)\boldsymbol{x}_j \right] \tag{14}$$

from which we deduce that

$$\sum_{i=1}^{N}\overline{\boldsymbol{x}_i} = \frac{(1-2\mu)N}{n_f+sn_o}\left[sn_o\mathbf{1}+\sum_{j=1}^{N}\overline{f(\boldsymbol{x}_j)\boldsymbol{x}_j}\right].$$

Let us now give, with less details, the derivation of the average value of $\sum f(\boldsymbol{x}_i)f(\boldsymbol{x}_j)\boldsymbol{x}_i\boldsymbol{x}_j$:

$$\begin{aligned}\left\langle\sum_{i=1}^{N}\sum_{j=1}^{N}f(\boldsymbol{x}_i')f(\boldsymbol{x}_j')\boldsymbol{x}_i'\boldsymbol{x}_j'\right\rangle &= \nu n_f + N(N-1)\\ &\times\left\{\sum_{i=1}^{N}\left[(1-\mu\nu)f(\boldsymbol{x}_i)\boldsymbol{x}_i+\mu\sum_{\alpha=1}^{\nu}f(M_\alpha\boldsymbol{x}_i)M_\alpha\boldsymbol{x}_i\right]\right.\\ &\left.\times\frac{(1+s\delta(\boldsymbol{x}_i-\mathbf{1})f(\boldsymbol{x}_i)}{n_f+sn_o}\right\}^2\\ &= \nu n_f+\frac{N-1}{N}\left(\frac{n_f-2\mu Np}{n_f+sn_o}\right)^2\\ &\times\left[s^2n_o^2\nu+2sn_o\mathbf{1}\sum_{i=1}^{N}f(\boldsymbol{x}_i)\boldsymbol{x}_i+\sum_{i=1}^{N}\sum_{j=1}^{N}f(\boldsymbol{x}_i)f(\boldsymbol{x}_j)\boldsymbol{x}_i\boldsymbol{x}_j\right].\end{aligned}$$

Using this result and equation (5), one gets:

$$\sum_{i=1}^{N}\sum_{j=1}^{N}\overline{f(\boldsymbol{x}_i)f(\boldsymbol{x}_j)\boldsymbol{x}_i\boldsymbol{x}_j} = \frac{\nu Nn_f+(N-1)\left(\frac{n_f-2\mu Np}{n_f+sn_o}\right)^2\left(1+2\frac{n_f-2\mu Np}{sn_o+2\mu Np}\right)s^2n_o^2\nu}{N-(N-1)\left(\frac{n_f-2\mu Np}{n_f+sn_o}\right)^2}.$$

On the same scheme, one evaluates the two remaining quantities:

$$\begin{aligned}\sum_{i=1}^{N}\sum_{j=1}^{N}\overline{f(\boldsymbol{x}_i)\boldsymbol{x}_j\boldsymbol{x}_j} &= \nu n_f+(N-1)\frac{(1-2\mu)(n_f-2\mu Np)}{(n_f+sn_o)^2}\\ &\times\left[s^2n_o^2\nu+2sn_o\mathbf{1}\sum_{i=1}^{N}\overline{f(\boldsymbol{x}_i)\boldsymbol{x}_i}+\sum_{i=1}^{N}\sum_{j=1}^{N}\overline{f(\boldsymbol{x}_i)f(\boldsymbol{x}_j)\boldsymbol{x}_i\boldsymbol{x}_j}\right]\\ \sum_{i=1}^{N}\sum_{j=1}^{N}\overline{\boldsymbol{x}_i\boldsymbol{x}_j} &= \nu N+N(N-1)\left(\frac{1-2\mu}{n_f+sn_o}\right)^2\\ &\times\left[s^2n_o^2\nu+2sn_o\mathbf{1}\sum_{i=1}^{N}\overline{f(\boldsymbol{x}_i)\boldsymbol{x}_i}+\sum_{i=1}^{N}\sum_{j=1}^{N}\overline{f(\boldsymbol{x}_i)f(\boldsymbol{x}_j)\boldsymbol{x}_i\boldsymbol{x}_j}\right].\end{aligned}$$

With these results, there is no difficulty in evaluating $\overline{\langle(\Delta\boldsymbol{X})^2\rangle}$ and $\overline{\sigma^2}$. The mean square shift is equal to the average of the expected value of the squared difference between the mean genotype $\boldsymbol{X}'$ of the offspring population and the mean genotype $\boldsymbol{X}$ of the parental one:

$$\overline{\langle(\Delta\boldsymbol{X})^2\rangle} = \overline{\langle(\boldsymbol{X}'-\boldsymbol{X})^2\rangle}.$$

By expanding the right-hand side of this expression, we have:

$$\overline{\langle(\Delta\boldsymbol{X})^2\rangle} = \frac{1}{N^2}\left[\sum_{i=1}^{N}\sum_{j=1}^{N}\overline{\langle\boldsymbol{x}_i'\boldsymbol{x}_j'\rangle}-2\sum_{i=1}^{N}\sum_{j=1}^{N}\overline{\langle\boldsymbol{x}_i'\rangle\boldsymbol{x}_j}+\sum_{i=1}^{N}\sum_{j=1}^{N}\overline{\boldsymbol{x}_i\boldsymbol{x}_j}\right].$$

By using

$$\sum_{i=1}^{N}\sum_{j=1}^{N}\langle \boldsymbol{x}_i'\boldsymbol{x}_j'\rangle = N\nu + \sum_{i\neq j}\langle \boldsymbol{x}_i'\rangle\langle \boldsymbol{x}_j'\rangle$$

and (14), we obtain an expression for $\overline{\langle(\Delta \boldsymbol{X})^2\rangle}$ in which all terms have been calculated above. Since the mathematical expression of this result is rather long, we do not give it explicitly here; however, in the absence of selective advantage ($s = 0$) and in the situation $p = 1$, $\overline{\langle(\Delta \boldsymbol{X})^2\rangle}$ is given by:

$$\overline{\langle(\Delta \boldsymbol{X})^2\rangle} = \frac{4\mu\nu}{N-(N-1)(1-2\mu)^2}. \tag{15}$$

The value of $\overline{\sigma^2}$ is calculated in the same way:

$$\begin{aligned}\overline{\sigma^2} &= \frac{1}{N}\sum_{i=1}^{N}\overline{(\boldsymbol{x}_i-\boldsymbol{X})^2}\\ &= \nu - \overline{\boldsymbol{X}^2}\\ &= \nu - \frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N}\overline{\boldsymbol{x}_i\boldsymbol{x}_j}.\end{aligned}$$

In the simple case $s = 0$ and $p = 1$, one finds:

$$\overline{\sigma^2} = (N-1)(1-\mu)\overline{\langle(\Delta \boldsymbol{X})^2\rangle}. \tag{16}$$

For an exhaustive study of this simple case see Higgs and Derrida (1991, 1992) and Derrida and Peliti (1991).

The results are presented in figure 1. It can be observed in figure 1(*b*) that the genomic variance remains tiny for $\mu < \mu_c$: the mean genotype remains confined in a small region of the sequence space $\mathcal{X}$, in the vicinity of the master genotype **1**; the genotypes present in the population are slightly dispersed around **1**. Moreover, the mutation rate being small, the mean square shift $\overline{\langle(\Delta \boldsymbol{X})^2\rangle}$ is also small (see figure 1(*a*)).

For $\mu > \mu_c$, the behaviour is very different. The dominant genotype **1** is no longer represented in the population; as a consequence, the genomic distribution broadens; however, the genomes remain localized in a small portion of the sequence space. Between two successive generations, there is an important random drift of the mean value $\boldsymbol{X}$ of the genotype in sequence space: the mean genotype wanders the sequence space $\mathcal{X}$. This is illustrated in figure 1(*c*). Figure 2 presents $\overline{\langle(\Delta \boldsymbol{X})^2\rangle}$ and $\overline{\sigma^2}$ versus the fraction p of fertile genotypes in $\mathcal{X}$.

The change of dynamical behaviour around μ_c is closely related to Eigen's error threshold. In Eigen's model, for $\mu < \mu_c$, the concentration of the various sequences are notably higher than zero only for sequences close to the fittest one (the master sequence) and the stationary distribution of frequencies is called the quasispecies; we have the very same behaviour in the present example. For $\mu > \mu_c$, Eigen's model predicts a homogeneous state: all sequences are equally represented in the system; our model predicts, for $\mu > \mu_c$, that the population remains grouped in the sequence space around the mean genotype $\boldsymbol{X}$ but that the latter one wanders the sequence space $\mathcal{X}$. In this respect, the present model corrects the unrealistic behaviour of Eigen's model mentioned in the introduction.

As can be seen in figure 1, the predicted curves fit the numerical data well, except in the transition region around μ_c. For $\mu \approx \mu_c$, the fluctuations of n_o can no longer be neglected as was done in the derivation of (13). This explains why the theoretically predicted threshold

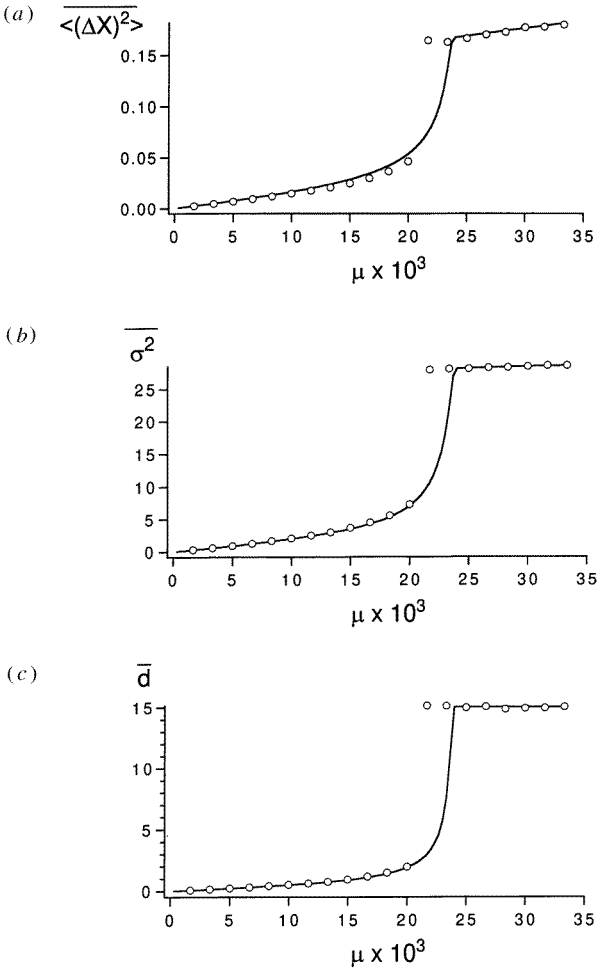

Figure 1. The evolution of $N = 200$ individuals is simulated. The genomes are built up with $\nu = 30$ monomers; each monomer has a probability μ to be badly copied during the duplication of the genome. The fraction of fertile genotypes is $p = 0.8$ and the genotype **1** has a selective advantage of $s = 2$. Graph (*a*) is a plot of $\overline{\langle(\Delta \boldsymbol{X})^2\rangle}$, the mean square shift of the mean genotype between two successive generations, versus μ. In (*b*), the average variance $\overline{\sigma^2}$ is plotted as a function of μ. (*c*) The average Hamming distance $\overline{d}$ between the mean genotype $\boldsymbol{X}$ and the favoured genotype **1** as a function of μ; for large values of μ, the average Hamming distance saturates at $\nu/2$. Dots correspond to numerical data, while the full curves represent the theoretical results.

overestimates μ_c. Another deviation between theoretical and numerical results can be seen in figure 2: for small values of p, the analytical prediction of $\overline{\langle(\Delta \boldsymbol{X})^2\rangle}$ becomes bad. This is mainly due to the following reason. Each genotype in $\mathcal{X}$ has ν nearest neighbours among which, on average, νp are fertile genotypes. If the average number of fertile neighbours becomes too small, fluctuations of the local density of fertile genotypes in $\mathcal{X}$—which are neglected in our 'mean-field' approach—begin to play an important role. It is, however, possible to improve the theoretical results for small values of p by specifying explicitly the distribution of fertile and sterile genotypes around the favoured genotype **1**. The next section is devoted to the study of this particular case.

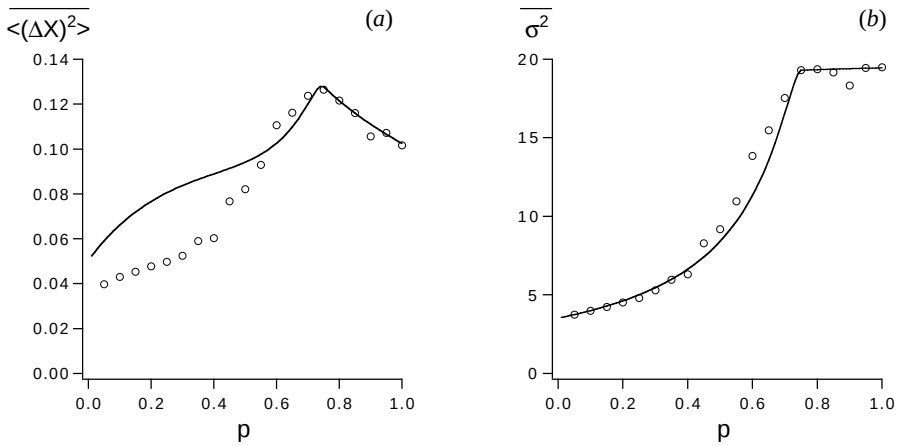

Figure 2. A system of $N = 200$ individuals with genomes of length $\nu = 20$ is simulated. The mutation rate is equal to $\mu = 0.0465$ and the genotype **1** has a selective advantage of $s = 10$. Graph (a) is a plot of the mean square shift of the mean genotype between two successive generations as a function of the fraction p of fertile genotypes in the sequence space $\mathcal{X}$. In (b), the variance is plotted versus p. Dots correspond to numerical data, while the continuous lines correspond to theoretical results. Due to the large value of s, the theoretical value for the critical mutation rate μ_c agrees fairly well with the one obtained numerically.

4. The small p limit

For small values of p, the quenched disorder is explicitly taken into account in the calculations. We consider, in sequence space $\mathcal{X}$, only the neighbourhood of the master genotype. The population evolves according to the following rules:

$$\boldsymbol{x}' = \begin{cases} \mathbf{1} & \frac{(1+s)n_o}{n_f+sn_o}(1-\mu\nu) \\ M_\alpha \mathbf{1} & \frac{(1+s)n_o}{n_f+sn_o}\mu \\ \boldsymbol{V}_\lambda & \frac{n_\lambda}{n_f+sn_o}(1-\mu\nu) \\ M_\alpha \boldsymbol{V}_\lambda & \frac{n_\lambda}{n_f+sn_o}\mu. \end{cases} \tag{17}$$

In this expression, $\boldsymbol{V}_\lambda$ for $\lambda = 1, \ldots, \Lambda \leqslant \nu$ are the genotype of the Λ first neighbours of **1** which are fertile. The population of $\boldsymbol{V}_\lambda$ is denoted by n_λ. The draw of another fertile sequence is not taken into account since it is assumed that its population is negligible. Evidently, it is possible to refine the model by considering them. We make the additional assumption that, for all λ, $\overline{n_\lambda} \equiv \overline{n}$. The study of this model follows the same lines as the preceding one. Thus, we begin by calculating the average quantities $\overline{n_o}$, $\overline{n_f}$ and $\overline{n}$ in the large t limit. Neglecting fluctuations in the relations $\overline{n_o} = \overline{\langle n_o \rangle}$, $\overline{n} = \overline{\langle n \rangle}$ and using the definition of n_f lead to the equations:

$$\begin{aligned} &\overline{n_o}(\overline{n_f} + s\overline{n_o}) = N[(1+s)(1-\mu\nu)\overline{n_o} + \mu\Lambda\overline{n}] \\ &\overline{n}(\overline{n_f} + s\overline{n_o}) = N[(1-\mu\nu)\overline{n} + (1+s)\mu\Lambda\overline{n_o}] \\ &\overline{n_f} = \overline{n_o} + \Lambda\overline{n}. \end{aligned} \tag{18}$$

Back-mutations on **1** are no longer neglected. These equations are easily solved in the large ν limit remembering that Λ is at most of order ν. The solution writes (for μ small enough):

$$\overline{n_f} = N(1 - \mu\nu + \Lambda\mu)$$

$$\overline{n_o} = N\left(1 - \mu\nu - \frac{1}{s}\Lambda\mu\right)$$
$$\overline{n} = N\frac{1+s}{s}\mu.$$

The mean fitness takes, on average, the same value $\overline{f_m} = (1+s)(1-\mu\nu)$ as in the former model. We neglect fluctuations of the populations in (17) by considering that they always take their mean values. The calculation of the quantities which deserve mention is straightforward since we have simply (no bar has been omitted):

$$\langle \boldsymbol{x}_i \rangle = \overline{\boldsymbol{x}_i} = (1-2\mu)\left(1 - 2\frac{\Lambda}{\nu}\frac{n}{n_f + sn_o}\right)\mathbf{1}. \tag{19}$$

The value of the moment $\overline{\boldsymbol{x}_i \boldsymbol{x}_j}$ is easily obtained:

$$\overline{\boldsymbol{x}_i \boldsymbol{x}_j} = \begin{cases} (1-2\mu)^2\left(1 - 2\dfrac{\Lambda}{\nu}\dfrac{n}{n_f + sn_o}\right)^2 & \text{if } i \neq j \\ \nu & \text{otherwise.} \end{cases}$$

Relation (19) between the average and expected values implies that the expression for $\overline{\langle(\Delta \boldsymbol{X})^2\rangle}$ simplifies to:

$$\overline{\langle(\Delta \boldsymbol{X})^2\rangle} = \frac{2}{N}(\nu - \overline{\boldsymbol{x}_1 \boldsymbol{x}_2}).$$

In the large ν limit, we obtain finally:

$$\overline{\langle(\Delta \boldsymbol{X})^2\rangle} = \frac{8\mu\nu}{N} + \frac{8\mu\Lambda}{Ns(1-\mu\nu)} \tag{20}$$

and

$$\overline{\sigma^2} = \frac{N-1}{N}\overline{\langle(\Delta \boldsymbol{X})^2\rangle}.$$

It remains to take the average over the disorder. For instance, it consists of replacing Λ with $p\nu$ in (20). The conclusion is that we can commute the average over the disorder with the other averages, but it is only true as $\nu \gg 1$. This validates the annealed disorder approximation in the former model. Numerical simulations confirm the validity of this approach, see figure 3 for an illustration.

5. Discussion

We have presented a method allowing us to calculate the time evolution of a population in the presence of selective advantage, sterile genotypes and mutations. This formalism can be viewed as an extension of Kimura's diffusion models. The population dynamics obtained with this approach are close to the one observed in Eigen's system of self-replicating macromolecules. There exists a critical value μ_c for the mutation rate μ; its value $\mu_c = s/\nu(s+p)$ depends essentially on two parameters: the selective advantage s of the favoured genotype and the fraction p of fertile genotypes in sequence space. For a small selective advantage s, the existence of sterile sites increases considerably the value of μ_c.

If $\mu < \mu_c$, the population remains confined in the neighbourhood of the genome having the highest selective advantage; if $\mu > \mu_c$, the population no longer remains fixed, but wanders the sequence space. In opposition to what happens in Eigen's system (where

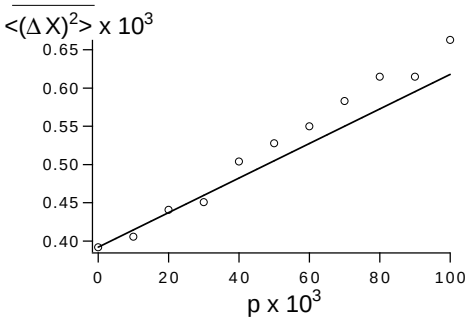

Figure 3. Comparison of numerical computations with theoretical results for a system with selective advantage and strong quenched disorder. The evolution of a population composed of $N = 10^4$ (chosen large enough in order to limit the effect of the fluctuations) individuals is simulated. The genomes are formed of $\nu = 25$ monomers. The genotype **1** has a selective advantage of $s = 0.2$. Each monomer has a probability $\mu = 0.02$ to be badly copied during the duplication of the genome. This graph is a plot of the average value of $\overline{\langle(\Delta \boldsymbol{X})^2\rangle}$ over many experiments as a function of p. Dots correspond to numerical data obtained by simulating a population obeying the rules defined in (17); the full curve corresponds to the theoretical result.

the system becomes homogeneous if $\mu > \mu_c$), the wandering population always keeps a structure: due to the reproduction mode, all individuals remain grouped in sequence space.

Finally, we have also shown that the study in the case of quenched disorder can be done as if we were in the presence of annealed disorder, at least as ν is large enough.

References

Derrida B and Peliti L 1991 Evolution in a flat fitness landscape *Bull. Math. Biol.* **53** 355–82

Eigen M 1971 Self-organization of matter and the evolution of biological molecules *Naturwissenschaften* **58** 465–523

——1993 Viral quasispecies *Sci. Am.* **269** 32–9

Eigen M, McCaskill J and Schuster P 1989 The molecular quasispecies *Adv. Chem. Phys.* **75** 149–263

Eigen M and Schuster P 1977 The hypercycle A: a principle of natural self-organization: emergence of the hypercycle *Naturwissenschaften* **64** 541–65

Higgs P G 1994 Error thresholds and stationary mutant distributions in multi-locus diploid genetics models *Gen. Res. (Camb.)* **63** 63–78

Higgs P G and Derrida B 1991 Stochastic models for species formation in evolving populations *J. Phys. A: Math. Gen.* **24** L985–91

——1992 Genetic distance and species formation in evolving populations *J. Mol. Evol.* **35** 454–65

Joyce G and Orgel L 1993 Prospects for understanding the origin of the RNA world *The RNA World* vol 1, ed R Gesterland and J Atkins (New York: Cold Spring Harbor Laboratory Press) pp 1–25

Kimura M 1962 On the probability of fixation of mutant genes in a population *Genetics* **47** 713–19

Maynard Smith J 1989 *Evolutionary Genetics* (Oxford: Oxford University Press)

Nowak M and Schuster P 1989 Error thresholds of replication in finite populations. Mutation frequencies and the onset of muller's ratchet *J Theor. Biol.* **137** 375–95

Peliti L and Bastolla U 1994 Collective adaptation in a statistical model of an evolving population *C. R. Acad. Sci., Paris (Life Sciences)* **317** 371–4

Prügel-Bennett A 1997 Modelling evolving populations *J. Theor. Biol.* **185** 81–95

Schuster P and Svetina J 1988 Stationary mutant distributions and evolutionary optimization *Bull. Math. Biol.* **50** 635–60

Serva M and Peliti L 1991 A statistical model of an evolving population with sexual reproduction *J. Phys. A: Math. Gen.* **24** L705–9

Woodcock G and Higgs P G 1996 Population evolution on a multiplicative single-peak fitness landscape *J. Theor. Biol.* **179** 61–73