J. Phys. A: Math. Gen. 24 (1991) L985-L991. Printed in the UK

LETTER TO THE EDITOR

Stochastic models for species formation in evolving populations

Paul G Higgs and Bernard Derrida

Service de Physique Théorique† de Saclay, CE-Saclay, F-91191 Gif-sur-Yvette Cedex, France

Received 17 June 1991, in final form 31 July 1991

Abstract. We introduce a simple model for sexual reproduction in a flat fitness landscape, the species formation model, which leads to a spontaneous division of a population into species. In this model mating between individuals is permitted only if their separation in genome space is less than a given value. This model shows non-self-averaging effects in the overlap distribution similar to what is observed in a one parent model representing asexual reproduction, and in contrast to a simpler model for sexual reproduction, the homogeneous population model, in which random pairing of individuals is permitted.

The theory of spin glasses, as developed in the past decade, led to an understanding of several features which characterize the spin-glass phase (at least at the mean field level): multivalley landscape, non-self-averaging effects, broad distribution $P(q)$ of the overlap etc (Mézard *et al* [1]). Although there is still a debate whether the mean field theory applies to real 3D spin glasses, the concepts orginally introduced for spin glasses are useful in other areas of physics (random interfaces, optimization etc) and biology (neural networks, protein folding, evolution etc). In this paper we discuss models of evolving populations with both sexual and asexual reproduction which seem to possess some features of the spin-glass phase.

Kauffman *et al* [2] have studied models for evolution in which the population is represented as a point evolving towards a local optimum in a rugged fitness landscape. Eigen *et al* [3] have studied populations of self-replicating macromolecules in which the replication rates of the sequences depend on the concentration of other sequences present, thus defining a complex fitness landscape which itself evolves in time.

Recently it was shown that even for a model of asexual reproduction in a flat fitness landscape (Derrida and Peliti [4]) the fluctuations in the number of descendants of each individual could lead to a non-trivial structure of overlaps in genome space. We refer to this as the one parent model (OPM). The most direct extension of this to sexual reproduction (Serva and Peliti [5]) allows random pairing between individuals. We will call this the homogeneous population model (HPM) since it leads to a uniform population, but in which pairing is only allowed between individuals closer than a certain distance in genome space. This model has non-self-averaging properties similar to those of the OPM.

In these three models each individual α is represented by a sequence of N Ising spins $\{S_1^\alpha, \ldots, S_N^\alpha\}$ which we call its genome. The number M of individuals at each generation is kept constant and this introduces a competition between the descendants of different individuals.

† Laboratoire de la Direction des Sciences de la Matière du Commissariat à l'Energie Atomique.

0305-4470/91/170985+07$03.50

In the OPM each individual at generation T has a parent chosen at random from the M individuals at generation $T-1$. The genome is inherited from the parent with small probability of errors governed by mutation rate μ. Thus if $G(\alpha)$ is the parent of α then

$$S_i^\alpha = S_i^{G(\alpha)} - \text{probability } \tfrac{1}{2}(1+e^{-2\mu}) = -S_i^{G(\alpha)} - \text{probability } \tfrac{1}{2}(1-e^{-2\mu}). \tag{1}$$

These probabilities result from a probability $\mu\,dt$ of mutation occurring in an infinitesimal time dt. Using an analogy with the random map model (Derrida and Bessis [6]) many properties of the genealogical tree and overlap distribution of the OPM can be calculated analytically [4].

The HPM is defined in a similar way [5]. Each individual α has two distinct parents $G_1(\alpha)$ and $G_2(\alpha)$ chosen at random from the previous generation. Each spin S_i^α is inherited from either $G_1(\alpha)$ or $G_2(\alpha)$ with equal probability, with the same probability of faithful copying or mutation as in equation (1).

A natural measure of the similarity of two individuals is their overlap

$$q^{\alpha\beta} = \frac{1}{N}\sum_{i=1}^{N} S_i^\alpha S_i^\beta. \tag{2}$$

In the limit $N \to \infty$ the models can be simulated directly by manipulating the overlap matrix $q^{\alpha\beta}$ rather than storing all the genome sequences.

In the OPM if the overlap between the parents $G(\alpha)$ and $G(\beta)$ of two individuals is $q^{G(\alpha)G(\beta)}$ then the expectation value of the overlap of α and β is

$$q^{\alpha\beta} = e^{-4\mu} q^{G(\alpha)G(\beta)}. \tag{3}$$

If N is infinite (3) becomes a deterministic rule for updating the overlap matrix, since fluctuations about the expectation value become negligible (of order $1/\sqrt{N}$). We thus simulate the OPM by choosing the set of parents $G(\alpha)$ at random and updating the matrix using (3). The diagonal elements $q^{\alpha\alpha}$ are kept equal to 1 always. If two individuals have a common ancestor $T^{\alpha\beta}$ generations ago then their overlap is $q^{\alpha\beta} = \exp(-4\mu T^{\alpha\beta})$. Thus there is a direct relation between the overlaps and the branching times on the genealogical tree.

There is an equivalent rule for updating the overlap matrix for the HPM in the limit $N \to \infty$. The pair of spins $S_i^\alpha S_i^\beta$ is inherited from one of the four combinations of parents of the two individuals with equal probability. Therefore

$$q^{\alpha\beta} = \frac{e^{-4\mu}}{4}\left(q^{G_1(\alpha)G_1(\beta)} + q^{G_2(\alpha)G_1(\beta)} + q^{G_1(\alpha)G_2(\beta)} + q^{G_2(\alpha)G_2(\beta)}\right). \tag{4}$$

Once again $q^{\alpha\alpha} = 1$ always.

The results of simulations of the OPM with population $M = 2000$ and mutation rate $\mu = 1/8000$ are shown in figure 1. Initially all the elements $q^{\alpha\beta} = 1$. The bottom curve shows the distribution $P(q)$ of the non-diagonal elements $q^{\alpha\beta}$ after an equilibration time of 4000 generations. Each subsequent curve (moving upwards) shows $P(q)$ 100 generations after the previous one. We see that $P(q)$ is non-self-averaging, i.e. it is not the same at each instant in time, and bears no resemblance to the time averaged distribution $\overline{P(q)}$. In fact in [4] it was shown that $\overline{P(q)} = \lambda q^{\lambda-1}$ where $\lambda = 1/4\mu M$. In this example $\lambda = 1$ and $\overline{P(q)}$ is a constant.

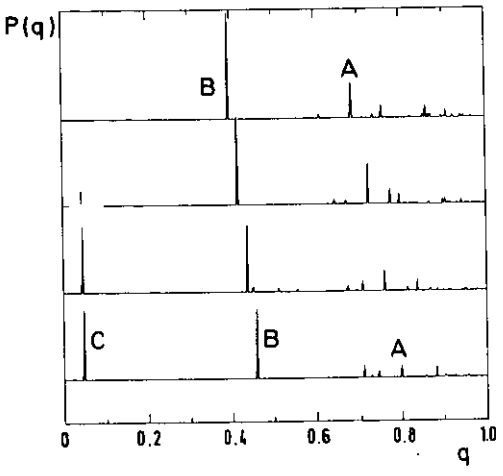

Figure 1. The overlap distribution $P(q)$ in the one parent model with $M = 2000$ and $\mu = 1/8000$ at four separate times. The earliest is at the bottom and subsequent curves (moving upwards) are at intervals of 100 generations. Peaks drift exponentially towards $q = 0$ (e.g. A and B). Some branches of the population (e.g. peak C) become extinct, whilst new peaks are forming constantly at $q = 1$.

Typically in figure 1 there are a large number of small peaks close to $q = 1$ and a small number of larger peaks at lower q values. This shows that the genealogical tree has many small branches at recent times which are descended from a few larger branches at earlier times. If we go back a time of order M generations in the past all the individuals are descended from the same ancestor.

For the HPM our results are shown in figure 2 for $M = 2000$ and $\mu = 1/8000$. $P(q)$ is shown at four times at intervals of 100 generations. There is essentially no difference between the curves: the model is self-averaging. For large M there is a probability $4/M$ that two individuals have one parent in common. Hence, to order $1/M$, the mean

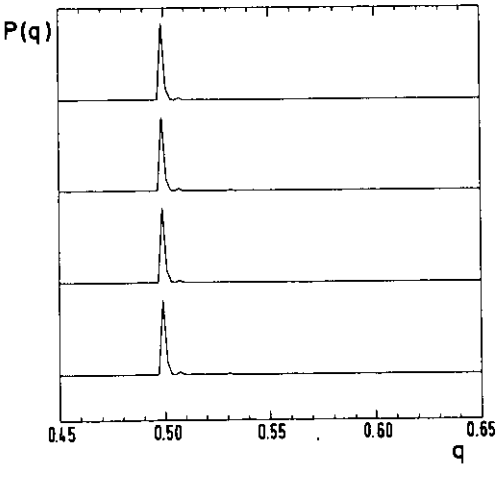

Figure 2. The overlap distribution in the homogeneous population model at four times, for $M = 2000$ and $\mu = 1/8000$. For large M the distribution becomes a single peak at the mean value $q = 1/2$, which does not fluctuate in time.

overlap at time T satisfies the following recursion

$$\langle q\rangle_{T+1} = e^{-4\mu}\left(\left(1-\frac{4}{M}\right)\langle q\rangle_T + \frac{4}{M}\left(\frac{3\langle q\rangle_T+1}{4}\right)\right). \tag{5}$$

If $\mu \ll 1$, the mean overlap tends towards the stationary value $q_0 = 1/(1+4\mu M)$. As calculated by Serva and Peliti [5], $P(q)$ has a single peak at q_0, which is $\frac{1}{2}$ in figure 2. Subsidiary peaks are seen at slightly higher q values, corresponding to pairs of individuals having an ancestor in common in a recent generation. These subsidiary peaks, and the variance of the distribution (which is also of order $1/M$) will disappear for large populations. This simplest version of a two-parent model therefore does not show the interesting properties of the OPM.

We now define the species formation model by introducing a parameter $q_{\min}$ and only allowing pairing between individuals with overlap $\geqslant q_{\min}$. This represents the fact that organisms too genetically different cannot produce a viable offspring. For each individual in the new generation the first parent $G_1(\alpha)$ is chosen at random. The second parent $G_2(\alpha)$ is then chosen at random from those individuals having $q^{G_1(\alpha)G_2(\alpha)} \geqslant q_{\min}$. If there is no second individual satisfying this requirement then a new first parent is chosen. The spins are then inherited from one or other of the parents, and mutations occur as before (equation (1)). In the long genome limit $N \to \infty$ the overlap matrix is updated as in (3).

We saw that all the overlaps in the HPM were close to a mean value q_0. The introduction of a $q_{\min}$ therefore only makes a difference if $q_{\min} \geqslant q_0$. Figures 3 and 4 show simulations of the SFM using the overlap matrix (limit $N \to \infty$) with $q_{\min} = 0.65$ and 0.9, again for $M = 2000$ and $\mu = 1/8000$, so that $q_0 = 1/2$. As in figure 1 $P(q)$ is shown at intervals of 100 generations with time moving upwards. Once again the distribution is a series of sharp peaks which move with time, indicating that the population has divided spontaneously into species.

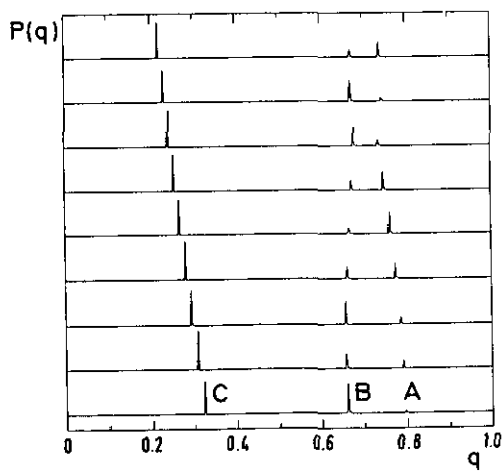

Figure 3. The overlap distribution for the species formation model with $M = 2000$, $\mu = 1/8000$ and $q_{\min} = 0.65$. $P(q)$ is shown at intervals of 100 generations (time moving upwards). Two species are present. Peaks A and B are the internal overlaps of members of the same species. These peaks wander randomly in the region $q > q_{\min}$. Peak C is the overlap between the two species, which drifts exponentially towards $q = 0$.

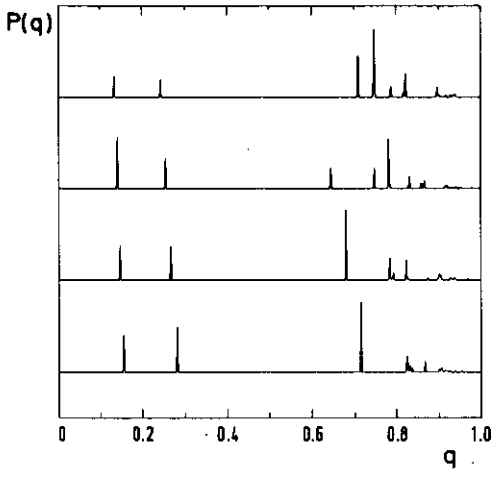

Figure 4. As figure 3 with $q_{min} = 0.9$. Many species are present, some of which die out, and some of which divide.

We think that this model can be understood qualitatively as follows. For the SFM, a species can be defined as a group of individuals having mutual overlaps greater than q_{min}. In figure 3 two species are present, yielding three main peaks. Each species behaves like a small version of the homogeneous population model. Because of the mixing of the characteristics between individuals within the species, all the overlaps between members of a species converge rapidly to a mean value, producing the two peaks A and B. If the population of species A is m_A then the mean q value of species A will drift slowly towards its natural value $q_0(m_A) = 1/(1+4\mu m_A)$. However, the population of each species changes with time (only the total $M = m_A + m_B$ is fixed), thus the mean value of q for individuals within a species appears not to be directly related to its population. Peak C represents the overlap between the two different species. The two species are not interbreeding, therefore the peak drifts exponentially towards $q = 0$. Peaks A and B seem to move rather randomly in the region $q > q_{min}$ due to stochastic fluctuations in the population sizes m_A and m_B. Eventually one of the species will die out leaving a single population of size M. The mean q for this species will then tend to drift below q_{min} since $q_0(M) \leqslant q_{min}$ and the population will once again have a tendency to split up.

According to this picture a species will only be stable if its population m is small enough such that $q_0(m) \geqslant q_{min}$. Hence the higher q_{min} the larger the number of species and the larger the number of peaks in $P(q)$. Figure 4 shows the case $q_{min} = 0.9$. Many peaks are seen which drift exponentially through the region $q < 0.9$. The result looks very similar to the one parent model in figure 1. Thus, in the SFM, large species tend to divide into smaller ones, which then either die out or grow into larger ones and divide again. This seems to mimic the behaviour of real living populations.

Many properties of the first two models have been calculated analytically [4, 5] because the evolution of the overlap matrix is linear. Introducing the q_{min} constraint makes the model highly nonlinear, and we have as yet been unable to find an analytical solution. Even quantities such as the mean overlap $\langle q \rangle$ and its time average $\overline{\langle q \rangle}$ can only be determined numerically. In figure 5 we show $\overline{\langle q \rangle}$ against q_{min}. For $M = 200$ the time average was taken over 40 000 generations, and for $M = 100$, over 10 000

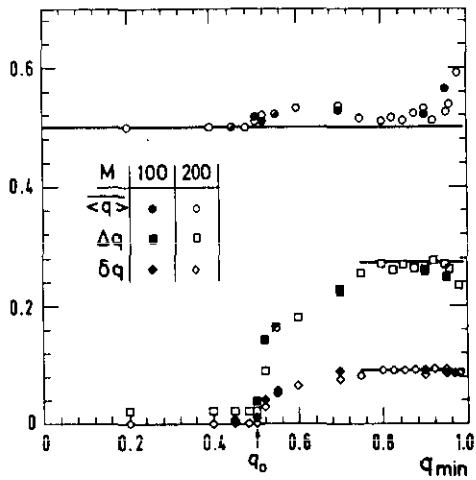

Figure 5. The mean overlap $\overline{\langle q \rangle}$ and its fluctuations δq and Δq for the SFM as a function of q_{min}. Population sizes 200 and 100 are shown, and μ is chosen such that $q_0 = 1/2$ in both cases.

generations. In both cases $\lambda(-1/4\mu m) = 1$ and therefore $q_0 = 1/2$. This allows direct comparison of the two population sizes. We see that $\overline{\langle q \rangle} = q_0$ for $q_{min} \leqslant q_0$, and is slightly higher when $q_{min} > q_0$. The rapid rise close to $q_{min} = 1$ is probably a finite size effect since the typical population of a species is very small in this limit if M is itself small. To investigate the self-averaging effects we also measured $(\delta q)^2 = \overline{\langle q \rangle^2} - (\overline{\langle q \rangle})^2$ and $(\Delta q)^2 = \overline{\langle q^2 \rangle} - \overline{\langle q \rangle^2}$, where $\langle \cdot \rangle$ represents an average within a population and the overbar represents a time average. The quantities δq and Δq are also shown in figure 5. For $q_{min} < q_0$ they should be negligible for large populations. For $q_{min} > q_0$ the system seems to be non-self-averaging since δq and Δq are of order 1 and do not seem to decrease with M. We appear to have a phase transition at $q_{min} = q_0$. The values of $\overline{\langle q \rangle}$, δq and Δq for the two populations are very similar which suggests that the SFM is governed largely by the parameter λ. The values of δq and Δq have been calculated in [4] for the OPM. We see in figure 5 that the measured values for the SFM seem to tend to the OPM values for large q_{min}.

In this work we have seen that the introduction of a minimum value q_{min} for the overlap of two individuals allowed to interbreed leads to a complex structure in genome space: the population breaks up into species. The overlap distribution within each species is homogeneous, whereas an interesting structure appears between the species, which is non-self-averaging.

One can think of many modifications to the SFM. First one could allow the size of the total population to fluctuate in time. Also one could replace the sharp cut-off $q^{G_1(\alpha)G_2(\alpha)} \geqslant q_{min}$ by a smoother condition, i.e. that once $G_1(\alpha)$ is chosen at random the probability of producing an offspring with another $G_2(\alpha)$ is equal to some smooth function $f(q^{G_1(\alpha)}, q^{G_2(\alpha)})$. Preliminary simulations suggest that these changes do not affect the qualitative picture observed above. There is, however, another change which has much more drastic effects. Suppose that, instead of choosing $G_1(\alpha)$ and then choosing $G_2(\alpha)$ from only those individuals satisfying the minimum overlap requirement, we choose a pair $G_1(\alpha)$ and $G_2(\alpha)$ at random. We then allow them to reproduce if $q^{G_1(\alpha)G_2(\alpha)} \geqslant q_{min}$, otherwise we choose a new pair. This second method would cause

any small species to disappear very quickly, because the size of the population $m_i(T+1)$ of species i at generation $T+1$ would be proportional to $m_i^2(T)$, whereas with the original method it is proportional to $m_i(T)$, which is more biologically reasonable.

From a biological point of view [7, 8] the models are clearly oversimplified. Assuming that genes are inherited independently neglects the linkage between genes on neighbouring parts of the same chromosome. Also we do not distinguish between the sexes of individuals (no male or female). Competition between species occurs only due to the finite population constraint, and there is no fitness landscape. Nevertheless, it is interesting that in spite of all these simplifications a population can produce species formation and a complex structure in genome space. The effect of a fitness landscape in the OPM has been investigated by Amitrano *et al* [9]. Introduction of a fitness landscape into the SFM would not alter the basic effect of division into species. Biological speciation is known to be greatly affected by geographical separation of sub-populations. We hope to investigate this in subsequent work. Our results suggest, however, that geographical separation is not a definite requirement for speciation.

From the point of view of a theoretical physicist the models are just simple rules for the evolution of the matrix $q^{\alpha\beta}$ (equations (3) and (4)) which produce complex structures in $P(q)$. In the OPM it is clear that the overlap distribution is ultrametric (Rammal *et al* [10]) due to the definition of the model. In the SFM the overlaps would also appear to be ultrametric since the peak widths are narrow. This feature has arisen spontaneously in this model. The rule for evolution of the matrix $q^{\alpha\beta}$ leads to a spontaneous symmetry breaking in a way similar to that observed in spin glasses [1] and automata models (Derrida and Flyvbjerg [11]). Inspired by this analogy we may develop various approximations to calculate the mean overlap and the typical number of species, etc, which we hope to present in future work. Whether one could find physical systems for which the overlap matrix evolves according to equations similar to (3) and (4) is still unclear to us.

We thank Luca Peliti for his encouraging comments.

References

[1] Mézard M, Parisi G and Virasoro M A 1987 *Spin Glass Theory and Beyond* (Singapore: World Scientific)
[2] Kauffman S A 1989 *Lectures in the Sciences of Complexity* ed D L Stein *(Proc. of the Summer School on Complex Systems, Santa Fe, 1988)* (Reading, MA: Addison-Wesley)
[3] Eigen M, McCaskill J and Schuster P 1988 *J. Phys. Chem.* **92** 6881
[4] Derrida B and Peliti L 1991 *Bull. Math. Biol.* **53** 355
[5] Serva M and Peliti L 1991 *J. Phys. A: Math. Gen.* **24** L705
[6] Derrida B and Bessis D 1988 *J. Phys. A: Math. Gen.* **21** L509
[7] Maynard Smith J 1989 *Evolutionary Genetics* (Oxford: Oxford University Press)
[8] Kimura M 1983 *The Neutral Theory of Molecular Evolution* (Cambridge: Cambridge University Press)
[9] Amitrano C, Peliti L and Saber M 1989 *J. Mol. Evol.* **29** 513
[10] Rammal R, Toulouse G and Virasoro M A 1986 *Rev. Mod. Phys.* **58** 765
[11] Derrida B and Flyvbjerg H 1987 *J. Physique* **48** 971; 1987 *J. Phys. A: Math. Gen.* **20** 5273