J Mol Evol (1992) 35:454–465

Journal of
Molecular Evolution

Genetic Distance and Species Formation in Evolving Populations

Paul G. Higgs and Bernard Derrida

Service de Physique Théorique, Centre d'Etudes de Saclay, F-91191 Gif-sur-Yvette Cedex, France

Summary. We compare the behavior of the genetic distance between individuals in evolving populations for three stochastic models.

In the first model reproduction is asexual and the distribution of genetic distances reflects the genealogical tree of the population. This distribution fluctuates greatly in time, even for very large populations.

In the second model reproduction is sexual with random mating allowed between any pair of individuals. In this case, the population becomes homogeneous and the genetic distance between pairs of individuals has small fluctuations which vanish in the limit of an infinitely large population.

In the third model reproduction is still sexual but instead of random mating, mating only occurs between individuals which are genetically similar to each other. In that case, the population splits spontaneously into species which are in reproductive isolation from one another and one observes a steady state with a continual appearance and extinction of species in the population. We discuss this model in relation to the biological theory of speciation and isolating mechanisms.

We also point out similarities between these three models of evolving populations and the theory of disordered systems in physics.

Key words: Genetic distance — Neutral theory — Speciation

Introduction

When studying living populations it is natural to try to classify individuals according to their similarities. To quantify these similarities it is useful to define a genetic distance between individuals which depends on the number of genes which they have in common. Individuals having common ancestors in the recent past will tend to have many genes in common: there will be a small genetic distance between them. At the level of species it is also possible to define measures of distance. These will normally depend on the time since two species descended from a common ancestral species. This can be done by observation of phenotypic characteristics (Sokal and Sneath 1963), or by direct observation of protein and DNA sequences (Goodman 1981; Felsenstein 1981; Bishop and Friday 1985).

In this article we discuss several models showing the behavior of genetic distances between individuals and between species. The models are stochastic and illustrate the importance of random fluctuations in gene frequencies in finite populations. The models are extremely simplified but we hope still retain some features relevant to real living populations. These models also illustrate the similarity between physical and biological systems.

We begin with the One-Parent Model (OPM), representing an asexually reproducing population (Derrida and Peliti 1991). In the next section we show how the genealogical tree is generated, and in the section after that we discuss the consequences of the tree structure for the behavior of the genetic distances.

Offprint requests to: P.G. Higgs

We compare the OPM to the Homogeneous-Population Model (HPM), which is a simple model for a sexually reproducing population (Serva and Peliti 1991). The continual mixing of genes in a random mating population causes the population to become homogeneous in the sense that the genetic distance between any pair of individuals is the same. For this model we shall see that genetic distances possess neither the tree structure or the large fluctuations observed in the OPM.

We then discuss the Species-Formation Model (SFM) (Higgs and Derrida 1991). In this model mating is possible only between individuals which are genetically similar to each other. We show that this model yields species which are in reproductive isolation from each other and that an evolutionary tree is again present if we look at the species level rather than the level of individuals.

Finally we discuss these models in connection with biology and theoretical physics, and give some possibilities for further research.

The models are defined in such a way as to be tractable to mathematical analysis in many cases (see Derrida and Peliti 1991; Serva and Peliti 1991). However, in this article we shall try to avoid mathematical details and illustrate the results mostly by numerical simulations.

The One-Parent Model

The OPM is defined as follows (Derrida and Peliti 1991; Higgs and Derrida 1991). The population consists of M individuals. Each individual α is represented by a sequence of N units: $\{S_1^\alpha, S_2^\alpha \ldots S_N^\alpha\}$. Here S_i^α is the i^{th} unit in sequence α. The sequence may represent the amino acids in a protein, or the bases of a nucleic acid sequence, or the alleles in a genome. We will refer to the units as alleles, and assume that each allele has two possible forms, so that each S_i^α can take the value $+1$ or -1.

A natural measure of genetic distance between individuals α and β is the Hamming distance:

$$d^{\alpha\beta} = \frac{1}{2}\sum_{i=1}^{N} | S_i^\alpha - S_i^\beta | \qquad (1)$$

This is just the number of alleles which are different in the two individuals. Another quantity containing the same information as $d^{\alpha\beta}$ is the overlap $q^{\alpha\beta}$, defined by

$$q^{\alpha\beta} = \frac{1}{N}\sum_{i=1}^{N} S_i^\alpha S_i^\beta = 1 - \frac{2d^{\alpha\beta}}{N} \qquad (2)$$

Two identical individuals have $d^{\alpha\beta} = 0$, and hence $q^{\alpha\beta} = 1$. If two genome sequences are completely independent, there is a probability ½ that two alleles S_i^α and S_i^β will be the same. Hence $d^{\alpha\beta} = N/2$ and $q^{\alpha\beta} = 0$.

We suppose that the size of the population M is fixed, and that each individual has an equal chance of producing offspring. Each individual α in one generation has a parent $G(\alpha)$ which is an individual chosen at random from the members of the previous generation. (This gives a Poisson distribution of the number of descendants of an individual, with some having many offspring and some having none. The mean number of offspring is of course 1.)

Each new individual inherits the genome of its parent, but with a small probability of error determined by the mutation rate μ. Thus

$$\begin{aligned} S_i^\alpha &= S_i^{G(\alpha)} \quad \text{with probability } \frac{1}{2}(1 + e^{-2\mu}) \\ S_i^\alpha &= -S_i^{G(\alpha)} \text{ with probability } \frac{1}{2}(1 - e^{-2\mu}) \end{aligned} \qquad (3)$$

For $\mu \ll 1$, these probabilities become $1 - \mu$, and μ, respectively. We assume that these mutations occur independently at different points on the sequence.

Before considering the behavior of the sequences themselves, and the overlaps between them, we will look at the hierarchical "family tree" generated by the model. Since each individual has a parent chosen at random from the previous generation, there is a probability $1/M$ that two individuals will have the same parent. Moreover, any two individuals can eventually be traced back to a common ancestor. The probability that the first common ancestor of individuals α and β occurred $T^{\alpha\beta}$ generations ago is $\overline{P}(T^{\alpha\beta})$, where

$$\overline{P}(T) = \frac{1}{M}\left(1 - \frac{1}{M}\right)^{T-1} \simeq \frac{1}{M}e^{-T/M} \text{ for large } M \qquad (4)$$

The bar indicates that $\overline{P}(T)$ is an average probability for all realizations of the family tree. If we look at the distribution of times $T^{\alpha\beta}$ between all pairs of individuals at one particular generation (Fig. 1), this bears no resemblance to the smooth function $\overline{P}(T)$.

Figure 1 was generated as follows. If the matrix $T_t^{\alpha\beta}$ is known at generation t then we can calculate it at the next generation simply by choosing randomly the M parents $G(\alpha)$ and using the relationship

$$T_{t+1}^{\alpha\beta} = T_t^{G(\alpha)G(\beta)} + 1 \qquad (5)$$

which relates the elements of the new matrix to the

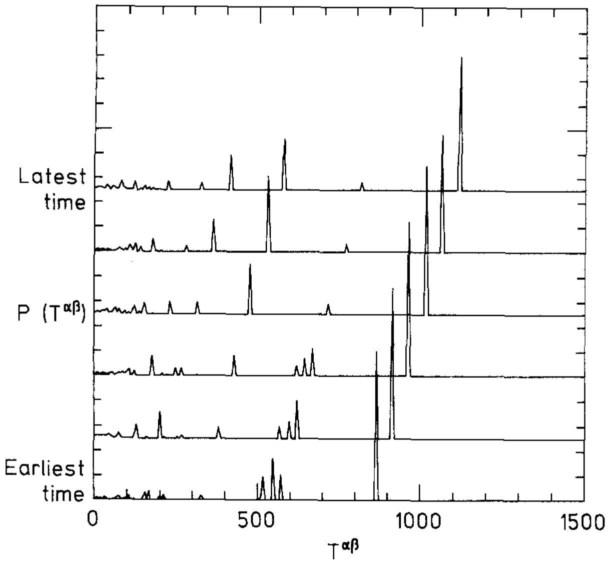


Fig. 1. Distribution of the elements of the matrix $T^{\alpha\beta}$ in the OPM for a population of $M = 1000$ individuals. The distribution is shown at six times for the same population. There is a period of 50 generations between each successive pair of curves; therefore the peaks move a distance 50 to the right each time. Peaks fluctuate in size and eventually disappear.

elements of the old matrix. At all times the diagonal elements $T^{\alpha\alpha}$ are kept equal to zero. We began with the initial conditions $T^{\alpha\beta} = 0$ for all α and β. The bottom curve of Fig. 1 shows the distribution $P(T^{\alpha\beta})$ of the elements of $T^{\alpha\beta}$ after a time of order M generations (in order to forget the initial conditions).

Each subsequent curve (moving upward) shows the distribution in the same population a short time (M/20 generations) afterward. A series of sharp peaks is seen, which can be understood in terms of the family tree of the individuals.

Figure 2 shows schematically the tree of descent of the current generation. $T^{\alpha\beta}$ is determined by the branch point representing the first common ancestor of α and β. There is one peak in Fig. 1 for each branch point of the tree. There are many small peaks at short times representing the large numbers of small branches at the top of the tree, while there are a small number of large peaks at longer times representing the small number of branch points at the base of the tree. As each new generation is added to the tree, the existing branch points move further back into the past; hence the peaks move steadily to the right in Fig. 1. New small peaks are constantly forming at $T = 1$. Any peak will eventually disappear due to random fluctuations in the number of descendants of the different individuals. The position of the earliest surviving branch point in the tree is typically a few times M generations in the past.

We can imagine cutting the tree at a particular point in the past (Fig. 2). This divides the population into a number of "families." The distribution of the sizes of these families can be calculated (Derrida and Peliti 1991) using the analogy with random map models (Derrida and Bessis 1988).

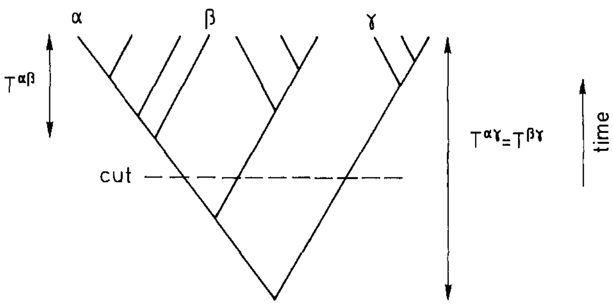


Fig. 2. Schematic representation of the genealogical tree in the OPM showing ultrametric property of the branching times $T^{\alpha\beta}$, $T^{\alpha\gamma}$, $T^{\beta\gamma}$. Cutting the tree at an arbitrary point in the past divides the population into families.

An important point about the matrix $T^{\alpha\beta}$ is that it is ultrametric; i.e., for any three individuals in the population

$$T^{\alpha\beta} \leq \max(T^{\alpha\gamma}, T^{\beta\gamma}) \qquad (6)$$

This has the consequence that the two largest of these three elements are equal. (For example $T^{\alpha\gamma} = T^{\beta\gamma}$ in Fig. 2.) For more details on ultrametric structures see Bishop and Friday (1985) and Rammal et al. (1986). The ultrametric structure in this model comes as no surprise since it is built into the model. However, we shall see that ultrametricity arises spontaneously in the Species-Formation Model to be discussed later.

Overlaps in the One-Parent Model

We will now look at the evolution in time of the overlap matrix $q^{\alpha\beta}$ defined in (2). Returning to the genome sequences of the individuals, we see that given the value of allele $S_i^{G(\alpha)}$ of the parent, the expectation value of that allele in the offspring is

$$E[S_i^{\alpha}] = e^{-2\mu} S_i^{G(\alpha)} \qquad (7)$$

which is calculated from the mutation probabilities in equation (3). Similarly, given the overlap between the parents $G(\alpha)$ and $G(\beta)$ of two individuals, the overlap between the offspring is

$$q^{\alpha\beta} = e^{-4\mu} q^{G(\alpha)G(\beta)} \qquad (8)$$

Strictly speaking, equation (8) applies only to the expectation value of $q^{\alpha\beta}$, but since $q^{\alpha\beta}$ is the sum of the contribution $S_i^{\alpha} S_i^{\beta}$ for many alleles i, and since mutations are assumed to occur independently in the different alleles, this equation becomes exact in

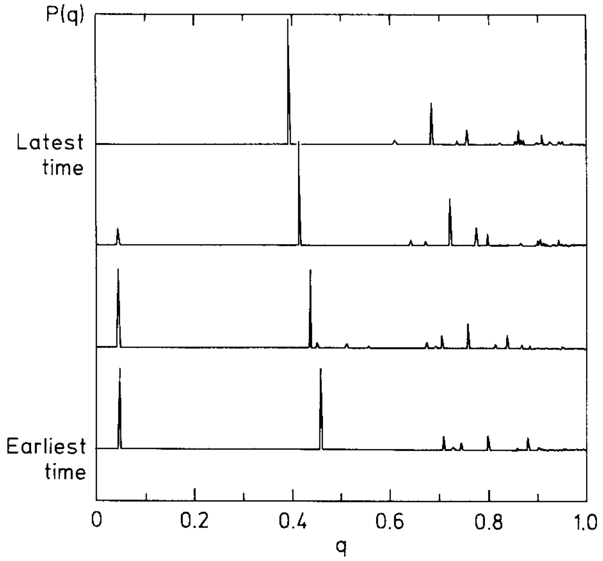


Fig. 3. Overlap distribution in the OPM, shown at four separate times separated by 100 generations for $M = 2000$, and $4\ \mu M = 1$. Peaks move exponentially toward $q = 0$.

the limit $N \to \infty$ (the infinite genome limit). Thus we can deal directly with the matrix of overlaps $q^{\alpha\beta}$ for the purpose of computer simulations (Higgs and Derrida 1991) and we do not need to store all the sequences.

We start with all individuals identical ($q^{\alpha\beta} = 1$ for all α and β). We then choose the parents $G(\alpha)$ randomly for each individual α in the new generation and create the new overlap matrix according to equation (8) for the nondiagonal elements $\alpha \neq \beta$. The diagonal elements $q^{\alpha\alpha}$ remain equal to 1 always. (The initial conditions are of course unimportant after a time of order M generations). This procedure is identical to the procedure for the time matrix $T^{\alpha\beta}$. In fact there is a direct relationship between the two quantities:

$$q^{\alpha\beta} = \exp(-4\mu T^{\alpha\beta}) \qquad (9)$$

The distribution of the elements of the $q^{\alpha\beta}$ matrix is shown in Fig. 3. Again there are many sharp peaks, and they drift exponentially toward $q = 0$, as the corresponding $T^{\alpha\beta}$ increases.

Thus $P(q^{\alpha\beta})$ contains a series of sharp peaks if measured at one moment in time, but when averaged over a long period in time the result is a smooth function $\bar{P}(q)$.

$$\bar{\mathrm{P}}(q) = \lambda q^{\lambda-1} \qquad \lambda = \frac{1}{4\ \mu M} \qquad (10)$$

This can be obtained simply by making the change of variables (9) in equation (4).

Note that the peaks in figure 3 are sharp simply because we have considered the long genome limit $N \to \infty$. For short sequence lengths, the peaks would be broadened to a width of order $1/\sqrt{N}$. If we wished to simulate the model for finite sequence length N we could not work directly with the overlap matrix as we do here, but we would have to store all the sequences and calculate explicitly all the mutations in each sequence. There would then be no strict link between $q^{\alpha\beta}$ and $T^{\alpha\beta}$ (equation 9), and the matrix $q^{\alpha\beta}$ would no longer be ultrametric. At any one time t, the population has an average overlap $\langle q\rangle_t$ which fluctuates in time about a mean value $\overline{\langle q\rangle}$. Here $\langle\ \rangle$ means an average over all individuals in one generation and the bar means a time average over many generations. From equation (8) we see that the time averaged mean overlap satisfies the equation

$$\overline{\langle \mathrm{q}\rangle} = e^{-4\mu}\left[\frac{1}{M} + \left(1 - \frac{1}{M}\right)\overline{\langle \mathrm{q}\rangle}\right] \qquad (11)$$

This is because there is a probability of $1/M$ that two individuals have a common ancestor, and hence the overlap of the parents in (8) is equal to 1. If $\mu \ll 1$ then the mean overlap has a solution which we call q_0

$$\overline{\langle q\rangle} = q_0 = \frac{1}{1 + 4\ \mu M} \qquad (12)$$

The mean value arises because of a balance between the mutations (tending to decrease q) and the common parentage factor (tending to increase q).

Population biologists often consider the inbreeding coefficient f, which is the probability that two randomly chosen homologous genes will be identical (Crow and Kimura 1970). The mean value $\bar{f}$ is calculated in a very similar way to equation (11). In fact

$$\overline{\langle q\rangle} = 2\bar{f} - 1 \qquad (13)$$

In a diploid population, the inbreeding coefficient is equivalent to the average homozygosity. The variance of f has been calculated by Stewart (1976) and Li and Nei (1975), and is shown to be large. The variance of $\langle q\rangle$, namely, $(\delta q)^2 = \overline{\langle q\rangle^2} - (\langle\bar{q}\rangle)^2$, is also large (Derrida and Peliti 1991). The two variances are related only indirectly, since q is an average property of all the loci on the sequence, whereas f is defined for one single locus.

Both $\overline{\langle q\rangle}$ and δq depend on the product μM, and are of the same order of magnitude. If we imagine taking the limit of large population size $M \to \infty$ in such a way that the product μM remains fixed, then δq will have a finite nonzero limit. Thus fluctuations about q_0 remain important even for large populations. For this reason we say that $\langle q\rangle$ is non-self-averaging, in contrast to what we will see in the HPM model.

The OPM which we have studied provides a way of visualizing the effects of random changes in genome frequencies in a finite population ("genetic drift"). Although some of the average quantities related to this model have long been known (Wright 1931; Crow and Kimura 1970), it is interesting to notice that properties of the population (such as $P(q)$) at a given instant in time may be very different from their average values, even for very large populations and very long genomes.

The Homogeneous-Population Model

The HPM is a simple model for a sexually reproducing population (Serva and Peliti 1991). As before, each individual is represented by a sequence of N alleles, each of which has two possible forms.

In this model, it is assumed that random pairing of individuals occurs, so each individual has two parents $G_1(\alpha)$ and $G_2(\alpha)$ randomly chosen from the previous generation. Each allele is inherited at random from one or other of the parents, (thus ignoring linkage between neighboring alleles). The allele is either a faithful copy from the parent or a mutation with the same probabilities as in equation (3).

As before, we know exactly the way the overlap matrix evolves in the limit $N \to \infty$.

$$q^{\alpha\beta} = \frac{e^{-4\mu}}{4}\left(q^{G_1(\alpha)G_1(\beta)} + q^{G_1(\alpha)G_2(\beta)} + q^{G_2(\alpha)G_1(\beta)} + q^{G_2(\alpha)G_2(\beta)}\right) \qquad (14)$$

This is because each pair of alleles $S_i^\alpha S_i^\beta$ contributing to $q^{\alpha\beta}$ comes with equal probability from one of the four possible combinations of the parents of α and β.

Figure 4 shows a simulation of the HPM of the same-size population and the same mutation rate as in Fig. 3. As before, only the overlap matrix was stored, not the genome sequences. The off-diagonal elements were updated according to equation (14) and the diagonal elements $q^{\alpha\alpha}$ remain equal to 1 always. We see that there is a single peak in $P(q)$ which remains stationary with time. From equation (14) the time-averaged mean overlap satisfies for large M

$$\overline{\langle q\rangle} = e^{-4\mu}\left\{\frac{4}{M}\left(\frac{3\overline{\langle q\rangle}+1}{4}\right) + \left(1-\frac{4}{M}\right)\overline{\langle q\rangle}\right\} \qquad (15)$$

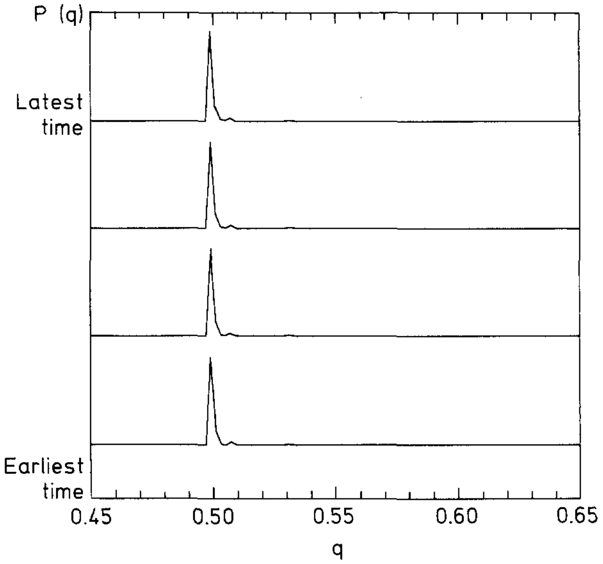


Fig. 4. Overlap distribution in the HPM for M = 2000, and $4\,\mu M = 1$. A large peak is seen at $q_0 = ½$ which does not move with time.

since for M ≫ 1 there is a probability $4/M$ that two individuals have a parent in common. (The probability that both parents are common is $O(1/M^2)$ and has been neglected.) The solution of equation (15) for $\mu \ll 1$ is the same mean value q_0 as in the OPM (equation 12).

The variance $(\delta q)^2$ of $\langle q\rangle$ has also been calculated by Serva and Peliti (1991). They found that δq vanishes in the limit $M \to \infty$, even if we impose the condition that μM is constant as we did in the OPM above. Thus $\langle q\rangle$ is self averaging in the HPM.

Examination of the data used to plot Fig. 4 reveals that there is not just one peak in $P(q)$, but there are several subsidiary peaks at slightly higher q values. These peaks are very small ($O(1/M)$) and are barely visible in Fig. 4. They represent the overlaps between individuals which have an ancestor in common in a recent generation. For example, if there was one grandparent in common, this would give an overlap $q = \frac{1}{16}(15q_0 + 1)$, hence the small peak at $q \simeq 0.53$. These subsidiary peaks are negligible for large populations.

We have called the population "homogeneous" because there is no family structure visible in $P(q)$. The genomes of the individuals may be thought of as a cloud of points in genome space with no organization into clusters as in the OPM. The fact that $P(q)$ does not change in time does not mean that the population is not evolving. In fact the cloud of points representing the population drifts randomly through genome space.

The overlap q_0 between any two individuals may be considerably less than 1 (depending on μM). Thus there is typically a large difference between any two individuals in the population.

The HPM lacks the species observed in real populations of sexually reproducing organisms. At the beginning of his book on species and evolution Mayr (1970, chapter 2) imagines a world without species, in which all individuals are members of a single random-mating population. Each mating pair would be widely different from each other, and from their descendants. This is precisely what happens in the HPM. We will now consider a model in which species do form.

The Species-Formation Model

The biological definition of a species is based on reproductive isolation (Mayr 1970). A species is a group of sexually reproducing organisms such that reproduction is possible between members of that group, but not between different groups. The SFM discussed is a model in which, due to the stochastic dynamics, species appear and disappear which are in reproductive isolation from one another.

The SFM is defined in the same way as the HPM earlier, except that rather than random pairing of individuals, pairing occurs preferentially between individuals which are genetically similar. We suppose that the first parent $G_1(\alpha)$ of individual α is chosen at random from the previous generation, but the second parent $G_2(\alpha)$ is chosen only from those individuals having an overlap $q^{G_1(\alpha)G_2(\alpha)}$ with the first parent greater than a cutoff value $q_{\min}$. Here, $q_{\min}$ is a parameter of the model which represents the presence of an isolating mechanism preventing reproduction between individuals which are too genetically different. Many types of isolating mechanisms are possible in biological systems (Mayr 1970; Maynard Smith 1989; Grant 1991). We discuss these mechanisms further in a later section.

Having chosen the parents of each new individual to satisfy the requirement that their overlap be greater than $q_{\min}$, we may then create the overlap matrix for the new generation according to equation (14). We know that in the absence of a cutoff there is a natural mean value of the overlap q_0 in the HPM. Therefore if $q_{\min} < q_0$ the cutoff makes no difference. On the contrary, for $q_{\min} > q_0$, the system is greatly perturbed by the cutoff since it can never reach its natural equilibrium state.

Figure 5 shows a simulation with $q_{\min} = 0.65$ and μ and M chosen so that $q_0 = ½$ as before. Once again several peaks are present which seem to move with time. $P(q)$ is non-self-averaging as in the OPM. If the cutoff is increased to $q_{\min} = 0.9$ (Fig. 6), then a larger number of peaks are present.

One can interpret these figures as follows. Figure 5 represents a situation where the population has split into two species. Peaks A and B represent the overlaps between members of the same species, and peak C represents the overlap between the two species. Since the species have overlap less than $q_{\min}$, no interbreeding is possible between them. Thus peak C moves exponentially with time toward $q = 0$ as the two species diverge. Each of the species behaves like a small independent version of the HPM. If species A has population m_A then it has a natural overlap $q_0(m_A) = 1/(1 + 4\mu m_A)$. (see equation 12). As long as m_A is not too large $q_0(m_A)$ will be greater than $q_{\min}$; and so breeding between members of the same species is not affected by the cutoff.

However, m_A fluctuates fairly rapidly from generation to generation. Only the total population M

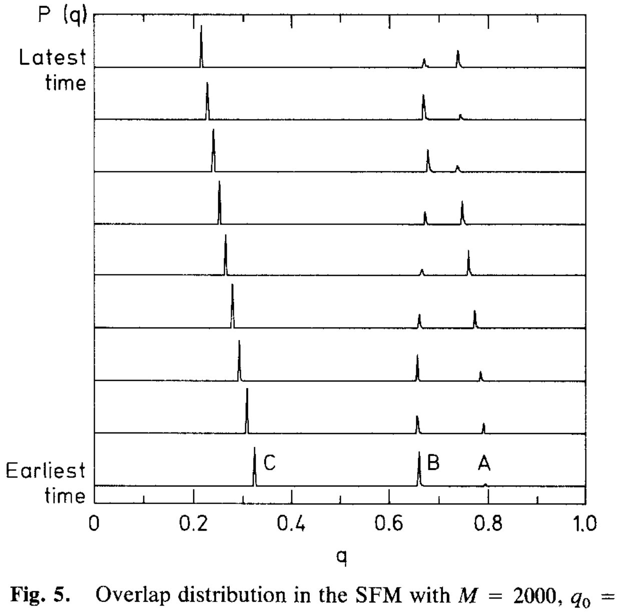


Fig. 5. Overlap distribution in the SFM with $M = 2000$, $q_0 = ½$, and $q_{\min} = 0.65$. This represents a situation where two species are present. (See text.)

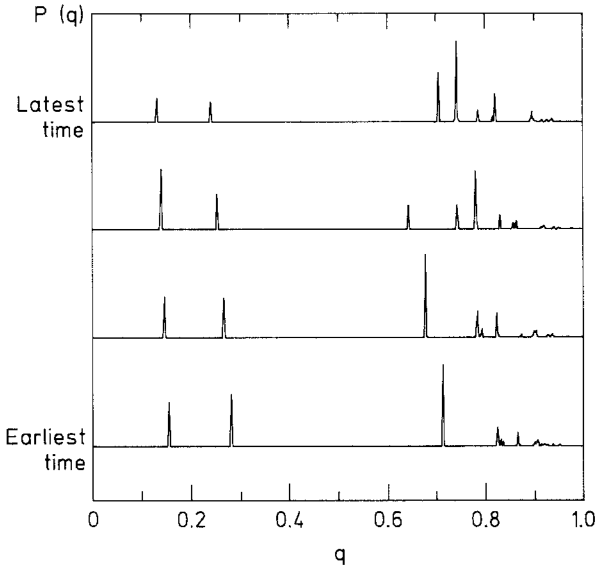


Fig. 6. Overlap distribution in the SFM, as in Fig. 5 except that $q_{\min} = 0.9$. Many species have formed and continual subdivision and extinction of species occurs.

of the whole system is fixed, not the population of each species. The movement of the mean overlap toward its natural value is rather slow since it is governed by the mutation rate. In other words, although the internal overlap is always tending toward $q_0(m_A)$, it never has chance to get there since $q_0(m_A)$ is itself changing due to fluctuations in m_A. The result is that peaks A and B move rather randomly in the range $q_{\min} < q < 1$ and there is no direct relationship between the weight of the peak (proportional to the square of the population size) and its q value. This is similar to the behavior of the homozygosity (or inbreeding coefficient) in a finite population (Kimura 1983; Maynard Smith 1989). In principle there should be a larger homozygosity in small populations than in large ones. However, it is difficult to observe this relationship since the population of a species is seldom sufficiently constant for these quantities to reach their equilibrium value. The homozygosity is particularly influenced by bottlenecks in the population size in the past.

The two-species situation in Fig. 5 is not stable in the long term. Eventually one or the other will die out due to fluctuations in the population sizes. Also, if by chance one species has a large population, its natural overlap will be less than $q_{\min}$. It will therefore tend to split into new species with smaller populations. Thus we have a continual appearance and disappearance of species.

We wish to note one important detail about the way the parents are chosen. The first parent $G_1(\alpha)$ is chosen at random. To select the second parent $G_2(\alpha)$, we continue to choose individuals at random until one is found having overlap greater than $q_{\min}$ with $G_1(\alpha)$. (If there is no such individual then $G_1(\alpha)$ is discarded. However, in practice this occurs very rarely.) The other possibility would have been to choose two individuals $G_1(\alpha)$ and $G_2(\alpha)$ at random and either accept or reject them both according to whether their overlap is greater than $q_{\min}$. Suppose there are $m_A(t)$ individuals in species A at time t. With the first method the expectation value in the next generation is $E[m_A(t + 1)] = m_A(t)$. With the second method it is $E[m_A(t + 1)] = M\ m_\alpha^2(t)/(\Sigma_k m_k^2(t))$, and thus small species would disappear very quickly. So only the first method correctly represents a neutral theory.

Overlaps Between Species

The sharp peaks which we see in Fig. 5 and 6 appear to indicate that we have well-defined species which are in reproductive isolation from one another. We will now show that by analyzing the matrix $q^{\alpha\beta}$ at any given time it is possible to assign each individual unambiguously to a species.

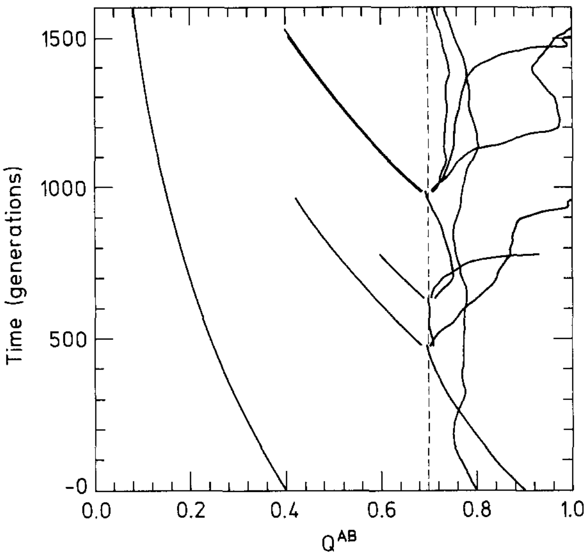


Fig. 7. The behavior of the elements of the species overlap matrix Q^{AB} as functions of time, for $M = 1000$, $q_0 = ½$, and $q_{\min} = 0.7$. Lines to the right of $q = 0.7$ represent internal overlaps of species. Lines to the left of $q = 0.7$ represent overlaps between species. Several speciation and extinction events are visible.

By definition we take the individual $\alpha = 1$ to be a member of species A. We then assign to species A any individual having an overlap greater than $q_{\min}$ with individual 1. Next we look for further individuals having an overlap greater than $q_{\min}$ with any individual in species A, and also assign them to species A. The process is repeated until there is no further individual which has an overlap greater than $q_{\min}$ with any of the members of A. Species A is then in reproductive isolation from all other individuals. We then look for the first individual in the list which is not a member of A, and this serves as a starting point for defining species B. The process is continued until every individual is assigned to a species.

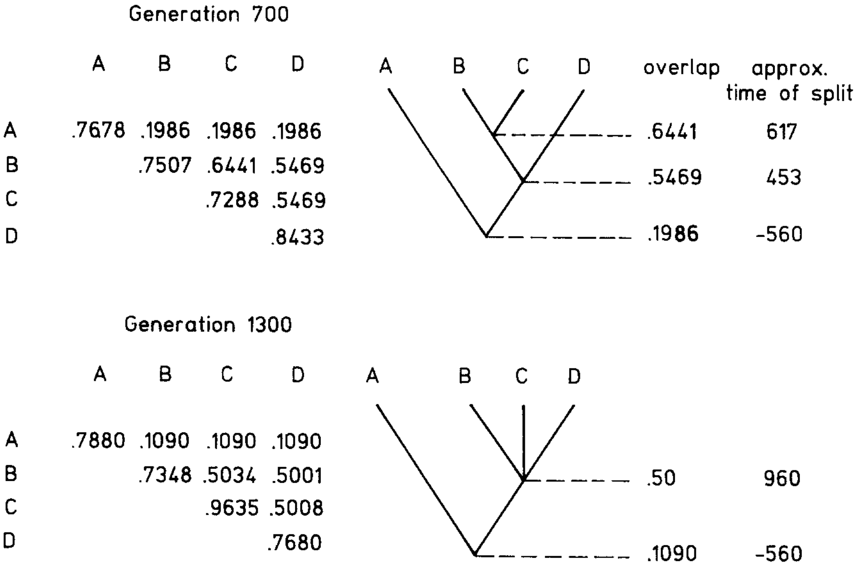
Suppose we find that there are K species and they have populations $m_A, m_B \ldots m_K$. It is possible to define a K by K matrix Q^{AB} which measures the similarity between species, rather than the M by M matrix $q^{\alpha\beta}$ which applies to single individuals. The elements of Q^{AB} are defined according to

$$Q^{AB} = \frac{1}{m_A m_B} \sum_{\alpha \in A} \sum_{\beta \in B} q^{\alpha\beta}$$

$$Q^{AA} = \frac{1}{m_A^2} \sum_{\alpha \in A} \sum_{\beta \in A} q^{\alpha\beta} \tag{16}$$

where "$\alpha \in A$" signifies that we take the sum over all individuals α belonging to species A.

Figure 7 shows the evolution in time of this species overlap matrix Q^{AB}. We have simply plotted a



Generation 700

	A	B	C	D
A	.7678	.1986	.1986	.1986
B		.7507	.6441	.5469
C			.7288	.5469
D				.8433

Generation 1300

	A	B	C	D
A	.7880	.1090	.1090	.1090
B		.7348	.5034	.5001
C			.9635	.5008
D				.7680

Fig. 8. Elements of the matrix Q^{AB} are shown at two times for the same example as in Fig. 7. The matrix is ultrametric to a good approximation and allows the genealogical tree of the species to be constructed. The approximate time of each speciation event can be calculated from the overlap value and these times are consistent with what we see in Fig. 7. The run was begun with two species present; therefore the time of the earliest split is apparently negative.

dot for each element of this matrix at each generation. The cutoff was $q_{\min} = 0.7$ in this example, and we began with two species with populations $M/2$ having internal overlaps 0.8 and 0.9, and an interspecies overlap of 0.4. Three speciation events are visible during the period of simulation. As expected this happens whenever one of the diagonal elements of Q comes close to $q_{\min} = 0.7$. For instance at time $\simeq 480$ we see the sudden disappearance of one line and the appearance of three new ones representing the internal overlap of two "daughter" species and the overlap between them. Several extinctions are also visible (e.g., $T \simeq 950$). If the population of a species goes to zero it must have passed through a period of small numbers, and hence the internal overlap will tend to be close to 1. An extinction is seen as the simultaneous disappearance of one of the lines representing an internal overlap, and one or more lines for the interspecies overlap.

In fact the matrix Q^{AB} is analogous to the similarity matrix between species which are obtained from comparison of real protein or nucleic acid sequences. In Fig. 8 we show the Q^{AB} matrix at times 700 and 1300 in the example of Fig. 7. From this data we can estimate the time T^{AB} since divergence of species A and B using the approximate relationship $Q^{AB} \simeq q_{\min}e^{-4\mu T^{AB}}$. The evolutionary trees constructed from these data can be compared with Fig. 7, which shows what actually happened in the simulation. The ultrametric inequality for overlaps implies that for any three elements Q^{AB}, Q^{BC}, and Q^{AC}, the two smallest must be equal (whereas the two largest times are equal). The example at time 700 is clearly an ultrametric matrix (down to four decimal places) and shows the presence of speciations occurring at times $\simeq 480$ and $\simeq 650$. The time-1300 example is more ambiguous because of the more complicated speciation event at time $\simeq 1000$. Here three new species have formed within a very short time, and the elements of Q show small deviations from ultrametricity, which we believe are due to the finite size of the population.

Such ambiguities are common in real data, and there is a large literature on methods of assigning the most likely tree to a given data set: Goodman (1981), Felsenstein (1981), Bishop and Friday (1985), and Blaisdell (1989). The problem is further complicated by insertions and deletions, so it is necessary to compare sequences which are of different lengths. Our model is much simpler in that it includes only point mutations. Sankoff and Kruskal (1983) discuss problems of sequence comparison in biology, physics, and computer science.

The method of assigning individuals to species adopted above ensures that no individual in one species has an overlap greater than $q_{\min}$ with any individual in a different species. However, it does not ensure that every pair of individuals assigned to the same species has an overlap greater than $q_{\min}$. We will now define a quantity σ_A which measures the "spread" of species A in genome space.

$$\sigma_A = \frac{1}{m_A^2} \sum_{\alpha \in A} \sum_{\beta \in A} \theta(q_{\min} - q^{\alpha\beta}) \qquad (17)$$

The step function $\theta(q_{\min} - q^{\alpha\beta})$ is 1 if $q^{\alpha\beta}$ is less than $q_{\min}$ and 0 otherwise. Thus the spread σ_A is simply the fraction of the elements $q^{\alpha\beta}$ between members of species A which are less than $q_{\min}$.

In Fig. 9 we show the behavior of σ_A with time together with the fraction $\phi_A = m_A/M$ of individuals belonging to species A and the internal overlap Q^{AA} of species A. We see that σ_A is identically zero for large periods of time and rises to high values over short periods corresponding to speciation events.

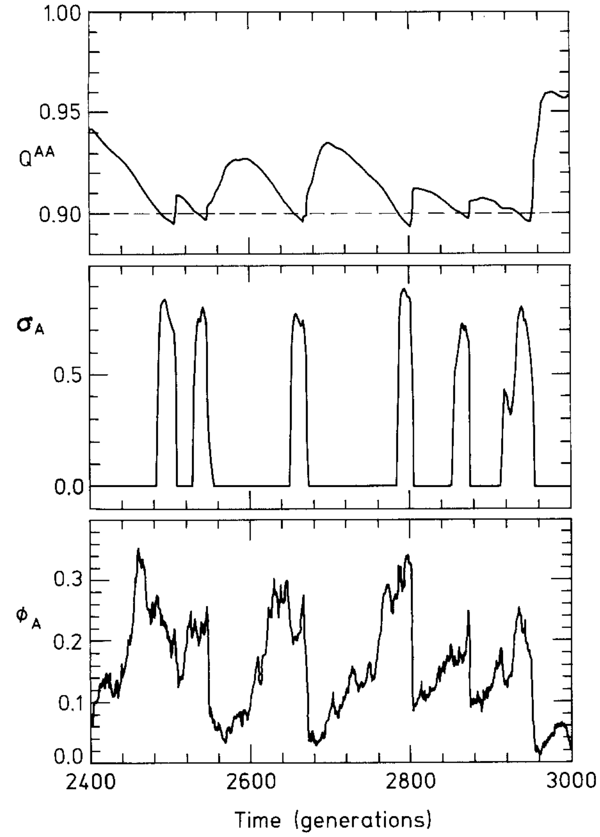


Fig. 9. The internal overlap Q^{AB} of species A is shown as a function of time for the SFM with $M = 1000$, $q_0 = ½$, and $q_{min} = 0.9$. When Q^{AA} drifts slightly below 0.9 speciation begins to occur. The "spread" σ_A defined in equation 17 is zero for long periods but becomes large at the moment of speciation. The fraction ϕ_A of individuals in species A is also correlated with Q^{AA} and σ_A. Speciation tends to occur when ϕ_A is large, and a sudden drop in ϕ_A occurs when the species divides.

The fraction ϕ_A follows a random walk. When ϕ_A becomes large the species begins to occupy a wide region of genome space and the spread σ_A becomes large. Division of A into separate species then occurs. This is seen as a sharp drop in ϕ_A during the periods in which σ_A is large. (To produce this picture we need to make sure that species A is always the same species at each generation. We do this by taking $G_1(1) = 1$ each time.)

Thus for most of the time $q^{\alpha\beta} > q_{min}$ for all pairs of individuals within a species (since $\sigma_A = 0$). Only at rare moments does σ_A become nonzero. This causes speciation to occur, and σ_A falls rapidly to zero again after the division. For this reason our method of assigning individuals to species appears rather reasonable.

Relationship of the Models to Biological Theory

The models above have been defined in a rather abstract way, and we wish to discuss some of the limitations and justifications of our models in terms of the biological theory of evolution and speciation.

First, we have at all points assumed that all individuals are equal irrespective of their genomes. We have done this because it is the simplest assumption. There is, however, good evidence that such a "neutral theory" is a good approximation at least for some cases of molecular evolution (Kimura 1983). Also, models similar to the OPM have been studied on various types of fitness landscape (Amitrano et al. 1989; Peliti 1990).

There have also been alternative approaches for studying the evolution of self-replicating macromolecules in rugged fitness landscapes (Kauffman and Levin 1987; Kauffman 1989; Rokhsar et al. 1986; Abbott 1988; Schuster and Swetina 1988; Fontana et al. 1989; Tarazona 1991). Here we have seen that there are many interesting effects observable even in a flat fitness landscape, and it would be interesting to know if the hierarchical structure of the population in the OPM and the speciation in the SFM will still be present in rugged landscapes. The analysis would, of course, be more complicated.

We have always assumed that parents of individuals are chosen randomly from the previous generation. This puts in the important feature that some individuals have no descendants, and others have one, or more than one. Using this model some individuals may have rather large numbers of offspring. We could avoid this by specifying a maximum number of offspring, for any one individual. This should not change any of the important features, but it may renormalize the time scales. (See Derrida and Peliti 1991.) The random choice of parents is convenient because the mathematical properties of the tree structure have already been studied in connection with random maps (Derrida and Bessis 1988; Derrida and Flyvbjerg 1987a).

Several modifications to the models for sexual reproduction can be envisaged which would make them more realistic. We could have distinct populations of males and females instead of allowing pairing between any two individuals. We could also include the fact that most sexual organisms are diploid by having two genome sequences within each individual. The important point of the HPM is, however, that the population becomes homogeneous. This would not be affected by the above modifications. More interesting would be to look at the effect of linkage between neighboring sites in the genome. This would be important if the number of chromosomes were very small, but the effect would be rather small for a typical species with say 20 pairs of chromosomes. We have also discussed the models in terms of a two-allele system ($S_i^\alpha = \pm 1$). Clearly one could generalize to any number of possible states for each S_i^α.

It is interesting to compare the family tree in the OPM with the tree of species generated by the SFM. In the OPM all individuals can be arranged directly on a tree because they have a single parent. In a random-mating population no such arrangement is possible since there are 2^T lines of descent leading to each individual stretching back T generations, and these lines of descent rapidly merge with those of all the other individuals. However, if we look at species rather than individuals, the species can be arranged on a tree, since any two current species presumably had an ancestral species in common at some point in the past. In the SFM, if we look at the level of the individual we have several independent species each behaving like a small version of the HPM. If we look at the species level then the tree structure becomes visible (by looking at the matrix Q^{AB}, for example). The term species really only has a meaning for sexually reproducing populations, since it depends on reproductive isolation. For asexual organisms the definition of species is largely a convention of the taxonomist. In our OPM we saw that by cutting the tree at a given point T in the past we create "families" such that $T^{\alpha\beta} \leqslant T$ (and hence $q^{\alpha\beta} \geqslant e^{-4\mu T}$) for all α and β in the family. Such families will be formed at whatever level of the tree we make the cut, and we could choose some arbitrary level to represent the species level if we wished.

We saw in the SFM that the introduction of a $q_{\min}$ leads to continuous process of division and extinction of species. The cutoff at $q_{\min}$ represents an isolating mechanism. There are many observed mechanisms which prevent the interbreeding between species (Mayr 1970; Grant 1991)—for example, anatomical differences, differences in courtship display, differences in flowering times in plants, and the inviability or infertility of hybrids. It is unlikely that these mechanisms are a strictly all-or-nothing affair like the sharp cutoff at $q_{\min}$ in our model. One could instead consider a smooth function $f(q)$ to represent the probability that successful mating occurs between individuals with overlap q. We have no idea what this function should be for a real organism, however. Any function $f(q)$ which is zero below a certain value $q_{\min}$ should give the same speciation phenomenon as the simple step function $f(q) = \theta(q_{\min} - q)$ which we used in the simulations above. We also tried a smooth sigmoid function $f(q) = 1/(1 + e^{\beta(q_{\min} - q)})$. This is nonzero even for very small q. The parameter β controls the sharpness of the cutoff. An instability in the distribution $P(q)$ was observed in this case representing splitting of the population into separate groups. Since there is no true reproductive isolation in this case the groups tend to merge back together again unless rather large values of β are chosen.

Isolating mechanisms may be divided into pre- and postmating barriers (Grant 1991). Postmating barriers such as the infertility or inviability of hybrids are likely to be rather general phenomena, while premating barriers are likely to be very specific mechanisms which prevent interbreeding between species which are very similar. If hybrid individuals are at a selective disadvantage then there will be a selection in favor of reproductive isolation (or reinforcement) (Mayr 1970; Maynard Smith 1989; Grant 1991). Individuals which mate preferentially with the same subspecies are then at an advantage. Hence the subspecies tend to diverge and become well-defined species. Crosby (1970) shows an interesting example of this happening between two plant subspecies. It may be possible to extend our model to distinguish more carefully between pre- and postmating barriers and to illustrate the reinforcement effect.

We have left out all effects of geographical isolation in our model. The speciation in the SFM is thus sympatric (occurring between individuals in the same location). There is much evidence (Mayr 1970) that naturally observed species have formed by allopatric speciation (i.e., in geographical isolation). Grant (1991) considers both sympatric and allopatric mechanisms to play a role. In any case sympatric speciation is at least a theoretical possibility (Maynard Smith 1966, 1989) and the SFM shows one way in which this could happen. It should be noted that the SFM has an inherent instability which leads to the initiation of the speciation process. We do not presuppose the existence of separate subspecies which are already different due to a heterogeneous environment (Maynard Smith 1966) or due to an initial period of geographical isolation (Crosby 1970).

Relationship of the Models to Disordered Systems in Physics

As physicists, our interest in these models of evolution began due to their similarities with spin glasses and other disordered systems (Peliti 1990). In the theory of disordered systems, it is often the case that phase space can be decomposed into several valleys of unequal sizes: in spin glasses, these valleys are free energy valleys or metastable states (Mézard et al. 1987; Binder and Young 1986); in random networks of automata or random map models (Derrida and Flyvbjerg 1987a,b; Fontanari 1991) they are the basins of attraction; and in protein folding models they would be the stable states of a protein (Shakhnovich and Gutin 1989).

The feature common to all these systems is that their phase space seems to be broken into different regions of random sizes which fluctuate from sample to sample in a similar way to that in which a randomly broken object gives rise to pieces of random sizes (Derrida and Flyvbjerg 1987b). If one performs the experiment of breaking dishes (an experiment rather easy to perform in anybody's kitchen), one knows that the size and the number of pieces will fluctuate from dish to dish with a few big pieces of random size and also many small pieces. The model, discussed above (OPM or SFM) with a population composed of several families (or species) is another example of a system broken into random pieces.

In certain spin glass models the overlaps between valleys have an ultrametric structure, as observed for the OPM and SFM. Rammal et al. (1986) discuss ultrametric structures as they occur in physics and biology. In spin glasses the overlap distribution $P(q)$ is known to be non-self-averaging at least at the mean field level. This means that it will be different if measured in two independent samples of large size. We saw that $P(q)$ in the OPM and the SFM was non-self-averaging. It is different if we look at two independent populations, or if we look at the same population at two widely separated moments in time. In the HPM, however, $P(q)$ is self-averaging: it is the same for all populations at all times in the limit of large size ($M \rightarrow \infty$). This is what usually happens in the thermodynamic limit in physics for most quantities studied, and probably also to $P(q)$ in spin glasses in low-enough dimension. Many people studying evolution on rugged fitness landscapes have used landscapes inspired by spin glass Hamiltonians (Amitrano et al. 1989; Kauffman 1989; Peliti 1990). The rugged landscape idea thus represents another point of similarity between the problems.

Several other recent articles in physics have a strong connection with evolutionary processes. Epstein and Ruelle (1989) have analyzed the numbers of species in the higher taxa of the plant classification system using models derived from branching processes in physics. Higgs and Orland (1991) have used a Monte Carlo method to simulate equilibrated ensembles of polymer configurations. The method is equivalent to the evolution of the ensemble in a rugged fitness landscape. Zhang et al. (1991) have looked at diffusion-reproduction processes in which the diffusion of points in real space is analogous to the diffusion of genome sequences in genome space in the OPM.

Conclusions

We have used the idea of overlaps to measure the similarity between genome sequences. In the One-Parent Model representing asexual reproduction the distribution of overlaps shows a series of sharp peaks reflecting the branching structure of the genealogical tree. The model provides an interesting way of seeing the consequences of the neutral theory of molecular evolution.

In the simplest model for sexual reproduction (HPM), with a random-mating population, the population becomes homogeneous, and no structure is seen in the overlap distribution.

If, instead of random mating, reproduction only occurs between individuals which are genetically similar we find that the population splits spontaneously into well-defined species. We can define a matrix of overlaps between these species which is approximately ultrametric and which has some analogy with the data obtained by comparing real protein and nucleic acid sequences. Analysis of this matrix allows a reconstruction of the history of the population.

The model suggests a method of sympatric speciation which may be relevant to real biological populations. It would be interesting to develop the model further to consider the relative importance of geographical and nongeographical effects in speciation, and to illustrate the selection in favor of reproductive isolation which would occur in certain cases.

Acknowledgments. We have benefited from discussions with W. Hamilton, S. Kauffman, and L. Peliti, and from correspondence with J. Maynard Smith.

References

Abbott LF (1988) A model of autocatalytic replication. J Mol Evol 27:114

Amitrano C, Peliti L, Saber M (1989) Population dynamics in a spin-glass model of chemical evolution. J Mol Evol 29:513

Binder K, Young AP (1986) Spin glasses: experimental facts, theoretical concepts and open questions. Rev Mod Phys 58: 801

Bishop MJ, Friday AE (1985) Evolutionary trees from nucleic acid and protein sequences. Proc Roy Soc Lond B226:271

Blaisdell BE (1989) Effectiveness of measures requiring and not requiring prior sequence alignment for estimating the dissimilarity of natural sequences. J Mol Evol 29:526

Crosby JL (1970) The evolution of genetic discontinuity: computer models of the selection of barriers to interbreeding between species. Heredity 25:253

Crow JF, Kimura M (1970) An introduction to population genetics theory. Harper and Row, New York

Derrida B, Bessis D (1988) Statistical properties of valleys in the annealed random map model. J Phys A Math Gen 21:L509

Derrida B, Flyvbjerg H (1987a) The random map model: a disordered system with deterministic dynamics. J Phys France 48:971

Derrida B, Flyvbjerg H (1987b) Statistical properties of randomly broken objects and of multi-valley structures in disordered systems. J Phys A Math Gen 20:5273

Derrida B, Peliti L (1991) Evolution in a flat fitness landscape. Bull Math Biol 53:355

Epstein H, Ruelle D (1989) Test of a probabilistic model of evolutionary success. Physics Reports 184:289

Felsenstein J (1981) Evolutionary trees from DNA sequences: a maximum likelihood approach. J Mol Evol 17:368

Fontana W, Schnabl W, Schuster P (1989) Physical aspects of evolutionary optimization and adaptation. Phys Rev A 40: 3301

Fontanari JF (1991) The adaptive map model. J Phys A Math Gen 24:L615

Goodman M (1981) Decoding the pattern of protein evolution. Prog Biophys Mol Biol 38:105

Grant V (1991) The evolutionary process. Columbia University Press, New York

Higgs PG, Derrida B (1991) Stochastic models for species formation in evolving population. J Phys A Math Gen 24:L985

Higgs PG, Orland H (1991) Scaling of polyelectrolytes and polyamphlytes—Simulation by an ensemble growth method. J Chem Phys 95:4506

Kauffman SA (1989) Lectures in the science of complexity. In Stein DL (ed) (Proceedings of the Summer School on Complex Systems, Santa Fe 1988). Addison-Wesley, Reading MA

Kauffman SA, Levin S (1987) Towards a general theory of adaptive walks in rugged fitness landscapes. J Theor Biol 128:11

Kimura M (1983) The neutral theory of molecular evolution. Cambridge University Press, Cambridge

Li WH, Nei M (1975) Drift variances of heterozygosity and genetic distance in transient states. Genet Res Camb 25:229

Maynard Smith J (1966) Sympatric speciation. American Naturalist 100:637

Maynard Smith J (1989) Evolutionary genetics. Oxford University Press, Oxford

Mayr E (1970) Populations, species and evolution. Harvard University Press, Cambridge

Mézard M, Parisi G, Virasoro MA (1987) Spin glass theory and beyond. World Scientific, Singapore

Peliti L (1990) A spin glass model of chemical evolution. Physica A 168:619

Rammal R, Toulouse G, Virasoro MA (1986) Ultrametricity for physicists. Rev Mod Phys 58:765

Rokhsar DS, Anderson PW, Stein DL (1986) Self-organization in prebiological systems: simulation of a model for the origin of genetic information. J Mol Evol 23:119

Sankoff D, Kruskal JB (1983) Time warps, string edits, and macromolecules: theory and practice of sequence comparison. Addison-Wesley, Reading MA

Schuster P, Swetina J (1988) Stationary mutant distributions and evolutionary optimization. Bull Math Biol 50:635

Serva M, Peliti L (1991) A statistical model of an evolving population with sexual reproduction. J Phys A Math Gen 24:L705

Shakhnovich EI, Gutin AM (1989) Formation of a unique structure in polypeptide chains. Theoretical investigation with the aid of a replica approach. Biophys Chem 34:187

Sokal RR, Sneath PHA (1963) Principles of numerical taxonomy. WH Freeman, San Francisco

Stewart FM (1976) Variability in the amount of heterozygosity maintained by neutral mutations. Theor Pop Biol 9:188

Tarazona P (1991) Error thresholds for molecular quasispecies as phase transitions: from simple landscapes to spin glass models. Preprint

Wright S (1931) Evolution in Mendelian populations. Genetics 16:97

Zhang YC, Serva M, Polikarpov M (1990) Diffusion reproduction processes. J Stat Phys 58:849

Received January 15, 1992/Revised April 24, 1992