Frequency distributions in population genetics parallel those in statistical physics

Paul G. Higgs
Department of Physics, University of Sheffield, Hounsfield Road, Sheffield S3 7RH, United Kingdom
(Received 27 July 1994)

A class of problems from statistical physics is discussed that is shown to be identical to a class of problems in population genetics. The mathematical treatment of these problems has arisen independently in the two subjects. The important results of both literatures are presented here, together with cross references. In each case there is a stochastic process generating a set of variables x_i that satisfy $\sum_i x_i = 1$. For example, the x_i may represent the weights of valleys in a spin glass, the sizes of attractors in dynamical systems, the frequency of different alleles in a population, or the sizes of different families in a genealogical tree. The frequency distributions $f(x)$ of the valleys or alleles are calculated, together with the distribution $\Pi(Y)$ of the quantity $Y = \sum_i x_i^2$. The distribution $\Pi(Y)$ can be written as a sum of universal functions $\Pi_k(Y)$ that are independent of the parameters of the problem. It is shown that the rather abstract concepts in the physical models are directly related to observables that are experimentally measureable in biology.

PACS number(s): 02.50.−r, 87.10.+e, 75.10.−b

I. INTRODUCTION

Much progress has been made recently in applying ideas taken from statistical physics to problems of evolution and mathematical biology. For example, the rugged fitness landscapes arising in evolutionary biology are sometimes modeled using rugged energy landscapes taken from spin-glass theory [1,2]; rugged energy landscapes are also relevant to the folding of biological macromolecules such as RNA [3,4]; branching tree structures and ultrametricity arise both in evolutionary biology and in spin-glass theory [5–8]; and the dynamics of coevolution in multispecies systems has recently been likened to the self-organized critical dynamics observed in some models of statistical physics [9,10]. In this article we will consider a class of problems where the parallel between physics and biology is extremely close. In fact, we will give two examples where the identical problem has arisen independently in the two fields. We will summarize the important results of the two literatures, so as to emphasize the similarities.

There are many problems in which one has a set of variables x_i that are determined by some stochastic process, such that $\sum_i x_i = 1$. One is interested in the probability distribution of these variables over many realizations of the process. For example, in disordered thermodynamic systems such as spin glasses [11–13], configuration space is divided into many low energy valleys separated by high energy barrier regions. In this case x_i represents the equilibrium probability of finding the system in the ith valley. Disordered systems contain quenched variables (the couplings J_{ij} in spin-glass models), which are usually determined from a random distribution. A different set of x_i arises from each random choice of the quenched variables. (Note that in most cases the symbol w_i is used for the weights of the valleys. We use x_i here since x is usually used for gene frequencies and w usually has a different meaning of fitness in the biological literature.)

A situation similar to the spin glass arises in dynamical systems such as neural networks and cellular automata [1,14,15]. Here phase space may have many attractors and x_i is the probability that the system falls into attractor i. Derrida and Flyvbjerg have studied two simpler problems of this type in which a great deal can be said analytically about the x_i distribution. One of these is the quenched random map [16], in which each configuration of the system has a randomly chosen successor configuration, resulting in a set of point attractors and limit cycles. The other example is the randomly broken object [17], in which an object to initial size 1 is broken by a random procedure into many pieces and x_i is the size of the ith piece.

Similar problems also arise in population genetics. Suppose that several different alleles for a given gene exist and x_i represents the frequency of the ith allele. These frequencies change in time due to random sampling, mutations, and selection. One may study the time averaged probability that the allele has frequency x_i [18–21]. The genealogical structure of the population is also of interest [22–25,6]. The population may be divided into families of closely related individuals and x_i then represents the fraction of the population in the ith family. The distribution of family sizes averaged over all realizations of the genealogical branching process may then be studied.

An interesting quantity in all these models is $Y = \sum_i x_i^2$. In genetics, Y is called the homozygosity [21,26,27]. It is the probability that two randomly sampled genes at a given locus are identical or that the two copies of a particular gene in a diploid individual are identical. In ecology, one may be interested in the relative abundances of different species within a community [28–30]. In this case Y is the probability that two randomly selected individuals are of the same species (this is sometimes called Simpson's index). In physics, Y arises in the replica theory of mean field spin glasses [11–13,15]. It is the

probability that two randomly chosen configurations fall in the same valley. If there are K valleys and all valleys have roughly equal occupation probabilities, then $x_i \sim 1/K$ and $Y \sim 1/K$. Thus in the thermodynamic limit, $K \to \infty$ and $Y = 0$ trivially. For spin glasses this happens at high temperature. The interesting thing about spin glasses is that there is a temperature T_c below which Y is nonzero in the thermodynamic limit. This is because there are a few valleys with weights x_i of order 1 and very many small valleys that do not contribute much to Y. Below T_c, Y is a random variable in the range 0–1 with a broad probability distribution $\Pi(Y)$. The same thing happens in population genetics: when there are many alleles present, a few of these have large frequencies, while most have very small frequencies. Thus the homozygosity Y has a broad and complex distribution.

II. ALLELE FREQUENCY DISTRIBUTIONS AND THE RANDOMLY BROKEN OBJECT

Let us begin with the K-allele model in population genetics (see Crow and Kimura [20], Chaps. 8 and 9, or Ewens [21], Chap. 3). Here one considers a gene that may exist in any one of K different forms (alleles). The population size is N and n_i is the number of copies of the ith allele in the population, so that the frequency of the ith allele is $x_i = n_i/N$. (We have assumed a haploid population for simplicity: for a diploid population it is necessary to replace N by $2N$ in all the following formulas.) In this model the alleles are neutral (i.e., they all have the same fitness), so that the gene frequencies at the next generation are determined by random sampling from the present generation. In addition, there is a mutation probability u per allele, and it is assumed that mutation occurs to one of the other $K-1$ alleles at random. This means that the probability that allele i is created by a mutation from a different allele is $v = u/(K-1)$. If the ith allele has frequency x_i at one generation, then the expectation value of the frequency at the next generation is $x_i' = x_i(1-u) + (1-x_i)v$. The number of copies n_i of this allele at the next generation has a binomial distribution

$$p(n_i) = (x_i')^{n_i}(1-x_i')^{N-n_i}\begin{bmatrix} N \\ n_i \end{bmatrix} . \tag{1}$$

If δx is the change in frequency from one generation to the next, one can obtain the mean and variance of δx from (1),

$$M_{\delta x} = -ux_i + v(1-x_i) , \tag{2a}$$

$$V_{\delta x} = x_i(1-x_i)/N , \tag{2b}$$

where small-order terms have been neglected in (2b). If we assume that $u \ll 1$ and $N \gg 1$, but the product uN is of order 1, then the discrete model discussed above can be approximated by a model in which time and frequency are continuous variables. This is usually called the diffusion approximation. Let $\phi(x,t;p)$ be the probability that the allele has frequency x at time t, given that it had frequency p at time zero. It has been shown ([20], Chap. 8; [21], Chap. 4) that ϕ satisfies the Kolmogorov forward equation

$$\frac{\partial \phi}{\partial t} = \frac{1}{2}\frac{\partial^2}{\partial x^2}(V_{\delta x}\phi) - \frac{\partial}{\partial x}(M_{\delta x}\phi) . \tag{3}$$

The discrete model discussed above (known as the Wright-Fisher model) is just one of a number of possible discrete models that have very similar diffusion approximations (see [21]). In the K-allele problem $\phi(x,t;p)$ converges to a stationary distribution $\phi(x)$ at long times, which is independent of p and is determined only by the parameter $\theta = 2Nu$ (or $4Nu$ for a diploid population),

$$\phi(x) = \frac{\Gamma(\theta K/(K-1))}{\Gamma(\theta)\Gamma(\theta/(K-1))} x^{\theta/(K-1)-1}(1-x)^{\theta-1} . \tag{4}$$

The limit $K \to \infty$ is of particular interest and is known as the infinite-alleles model. In this case each new mutation creates an allele that has never before existed in the population. There is no true stationary distribution for the frequency of any given allele x_i since each allele is bound to go to extinction eventually. Thus the limit of (4) when $K \to \infty$ is not well defined. However, the function

$$f(x) = \lim_{K\to\infty}[K\phi(x)] = \theta x^{-1}(1-x)^{\theta-1} \tag{5}$$

is well defined and is known as the frequency spectrum. The mean number of alleles in the population with frequency between x and $x+dx$ is $f(x)dx$. Since $f(x)$ diverges like x^{-1} for small x, there is a large number of very low frequency alleles. The function $f(x)$ is not normalizable. This is because we have taken the limit $K \to \infty$ and $N \to \infty$ with constant θ, so that the number of alleles is infinite. The function $g(x) = xf(x)$ is normalized so that $\int_0^1 g(x)dx = 1$. This has the interpretation that if we take a random sample from the population, the probability that the allele obtained has frequency between x and $x+dx$ is $g(x)dx$.

We will now consider the problem of the randomly broken object introduced by Derrida and Flyvbjerg [17] and show that this is identical to the finite-alleles model. Starting with an object of size 1, a fraction ξ_1 is broken off, giving a piece of size x_1, which is retained. From the remainder a fraction ξ_2 is broken off to give a piece of size x_2 and a further remainder, which is again broken, and so on. Repeating this process gives an infinite set of pieces with sizes

$$\begin{aligned} x_1 &= \xi_1 , \\ x_2 &= (1-\xi_1)\xi_2 , \\ x_3 &= (1-\xi_1)(1-\xi_2)\xi_3 , \end{aligned} \tag{6}$$

etc. The ξ_i are independent random variables in the range 0 to 1 chosen according to a given probability distribution $\rho(\xi)$. Following [17] we let $g(x) = \overline{\sum_i x_i \delta(x-x_i)}$, where the overbar indicates an average over all realizations of the breaking process, i.e., over all random choices of the ξ_i for a given function $\rho(\xi)$. It has been shown that $g(x)$ satisfies the integral equation

$$g(x) = x\rho(x) + \int \rho(\xi) g(x/(1-\xi))d\xi . \tag{7}$$

Derrida and Flyvbjerg then chose a particular functional form $\rho(\xi)=\theta(1-\xi)^{\theta-1}$, which makes the solution of (7) simple. The solution is $g(x)=xf(x)=\theta(1-x)^{\theta-1}$, which is identical to (5) obtained for the infinite-alleles model. In this case $f(x)$ is the mean number of pieces formed of size x. To prove that the models are really equivalent, it is necessary to look at the joint distribution of sizes of two or more pieces. For example, it can be shown that the mean number of pairs of pieces (alleles) in the same sample with sizes (frequencies) x and x' is $f(x,x')=\theta^2(1-x-x')/xx'$ in both models and that the higher-order joint distributions are also identical. It is a coincidence that the choice of $\rho(\xi)$ made in [17] corresponds to the one that arises naturally in the infinite-alleles model. The procedure of Eq. (6) has also been used by Patil and Taillie [30] (who refer to it as "preemption") and by Donnelly [31]. The list of pieces formed by this procedure is a "size-biased permutation." The largest pieces tend to occur at the beginning of the list, although they are not strictly ranked in decreasing order of size. The analogy between these frequency distributions and a broken object has previously been made by MacArthur [28], who discusses a "broken stick" model for the distribution of frequencies of bird species.

Several of the other problems mentioned above have similar $f(x)$ distributions. The quenched random map [16] is the same as Eq. (5) with $\theta=\frac{1}{2}$. For the mean field spin glass $f(x)=\text{const}x^{y-2}(1-x)^{-y}$, where $y=1-T/T_c$. This is the same as the K-allele model [Eq. (4)] for suitable choice of parameters; however, the problems are not identical since there are only K alleles in this case, while there is an infinite number of valleys in the spin glass. The joint distributions of the type $f(x,x')$ are not equivalent for these problems.

We will now consider the homozygosity Y in the infinite-alleles model. The mean and the variance of Y can be obtained as

$$\overline{Y}=\int dx\, x^2 f(x)=1/(1+\theta)\,, \tag{8}$$

$$\overline{Y^2}=\int dx\, x^4 f(x)+\int dx \int dx'(xx')^2 f(x,x')$$

$$=6/(1+\theta)(2+\theta)(3+\theta)\,, \tag{9}$$

$$\text{var}(Y)=\overline{Y^2}-(\overline{Y})^2=2\theta/(1+\theta)^2(2+\theta)(3+\theta)\,. \tag{10}$$

In the biological literature this variance was obtained by Stewart [26]. We are normally interested in large populations N with small mutation rates u, so that θ may be of order 1. In this case the fluctuations in Y are of the same order as the mean and hence Y is non-self-averaging. We have previously shown that there are many other quantities related to neutral evolution that are non-self-averaging (Higgs and Derrida [7,8] and Higgs and Woodcock [32]). The important consequence of this is that there are very large fluctuations between different realizations of the same evolutionary process.

The homozygosity Y is important in population genetics since quantities related to Y have been used as tests for the neutral theory of evolution (Watterson [27] and Tavaré, Ewens, and Joyce [33]). The case of three alleles only ($K=3$) has been studied analytically (Stewart [26]), and the general K case has been studied by simulation. One way to generate $\Pi(Y)$ is by simulation of the random sampling and mutation procedure. We have done this for the infinite-alleles model using a population of $N=1000$ and four different u values chosen to give $\theta=\frac{1}{4}$, $\frac{1}{2}$, 1, and 2. Figure 1 shows the histograms for $\Pi(Y)$ obtained after 10^6 generations. This requires considerable computer

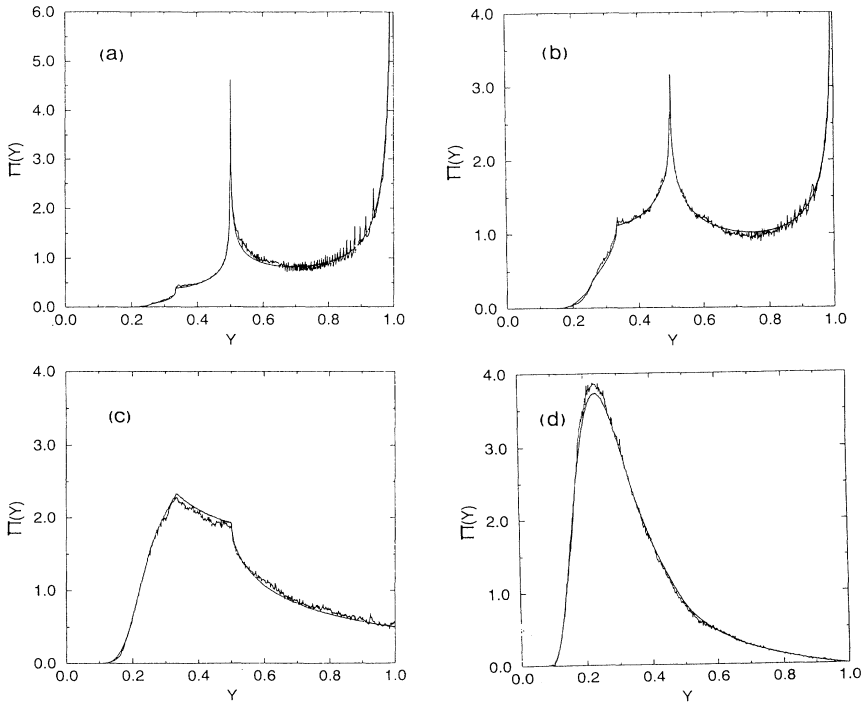

FIG. 1. The distribution of the homozygosity in the infinite alleles model is identical to the $\Pi(Y)$ distribution for the randomly broken object. The curves are calculated for four values of θ: (a) $\frac{1}{4}$, (b) $\frac{1}{2}$, (c) 1, and (d) 2. The smooth curves are calculated by a rapid iterative procedure given originally for the randomly broken object, while the noisy curves are calculated by explicit simulation of the infinite alleles model. (b) is also the solution of the quenched random map problem.

time since many random variables need to be generated to create each new generation and only one new Y value is created each generation. Since we have shown that the infinite-alleles model is identical to the randomly broken object, we can use a trick given in [17] to generate an accurate $\Pi(Y)$ distribution very rapidly. Suppose we have a sample of pieces x_i and a corresponding value of Y. We can add another piece x with probability distribution $\rho(x)=\theta(1-x)^{\theta-1}$ and simultaneously shrink all the other pieces by a factor $(1-x)$. We now have a new sample with a new Y value given by $Y'=x^2+(1-x)^2Y$. Iterating this formula requires only one random variable for each new Y value created and hence is much quicker than simulating the infinite-alleles model. [If r is a random number with a uniform distribution, then $x=1-r^{1/\theta}$ has the required distribution $\rho(x)$.] The smooth curves in Fig. 1 were generated using 10^9 iterations of this procedure. These figures indicate that the same function can be generated using two entirely different methods.

$\Pi(Y)$ behaves like $(1-Y)^{\theta-1}$ close to $Y=1$; thus there is a divergence as $Y\to 1$ if $\theta<1$. There are also singularities in $\Pi(Y)$ at each simple fraction $Y=1/K$. These are clearly visible in Fig. 1 for $Y=\frac{1}{2}$ and $\frac{1}{3}$ and become less pronounced for larger K. They are also less pronounced for larger values of θ, but are still present nevertheless. The singularities can be understood in the following way. In the infinite-alleles model the number of distinct alleles actually present in the population fluctuates. If there are K alleles present, then the minimum value of Y is $1/K$, when all alleles have equal frequency. If $Y<1/K$, then there must be more than K alleles present. There is thus an extra constraint added to the frequency distribution every time Y crosses a fraction $1/K$, so that $\Pi(Y)$ has a different functional form in each interval $1/K$ to $1/(K+1)$.

III. GENEALOGIES AND RANDOM MAPS

We now turn to the structure of genealogical trees in neutral evolution. It has been shown by Derrida and Peliti [6] that the annealed random map [34] is identical to the Wright-Fisher model of population genetics [18–21]. Each of the N individuals in the population has a parent chosen randomly from the previous generation. Following this procedure back in time creates a family tree which eventually leads back to a single common ancestor. The time sale is proportional to N, so it is convenient to introduce a scaled time $\tau=$(time in generations)$/N$. If one considers any given time τ in the past, individuals may be grouped into mutually exclusive τ families such that all pairs of individuals within a family have their latest common ancestor less than τN generations ago. Derrida and Bessis [34] give the distribution of sizes of the τ families. We will derive their result in a different way, using the diffusion approximation, which is standard in population biology.

Consider a particular individual at time 0, let x be the fraction of the population at time t which is descended from this individual, and let $\phi(x,t;p)$ be the probability distribution of x. The initial condition is $x=p=1/N$ for a haploid population of size N. Equation (3) applies with $M_{\delta x}=0$ and $V_{\delta x}$ as in (2b). An exact solution is possible in terms of an eigenfunction expansion (Crow and Kimura [20], Sec. 8.4)

$$\phi(x,t;p)=\sum_{i=1}^{\infty}\frac{(2i+1)(1-r^2)}{i(i+1)}T_{i-1}^1(r)T_{i-1}^1(z)\times\exp[-i(i+1)t/2N]\,, \tag{11}$$

where $z=1-2x$ and $r=1-2p$. The Gegenbauer polynomials $T_i^1(z)$ satisfy the recursion

$$(i+2)T_{i+2}^1(z)=(2i+5)zT_{i+1}^1(z)-(i+3)T_i^1(z)\,, \tag{12}$$

with $T_0^1(z)=1$ and $T_1^1(z)=3z$. We are interested in the mean number of τ families of size x, which is $f_\tau(x)=N\phi(x,N\tau;1/N)$ when $N\gg 1$:

$$\begin{aligned} f_\tau(x)&=Z_1(\tau)\delta(x-1)\\ &\quad+\tfrac{1}{2}\sum_{i=1}^{\infty}(2i+1)T_{i-1}^1(z)\exp[-i(i+1)\tau/2]\\ &=Z_1(\tau)\delta(x-1)+\sum_{k=2}^{\infty}Z_k(\tau)k\phi_k(x)\,, \end{aligned} \tag{13}$$

where $Z_k(\tau)$ is the probability that there are k families (which depends on τ) and $\phi_k(x)$ is the probability that a family has size x given that there are k families (which does not depend on τ). The terms in the sum in the first line of (13) arise directly from (11), while the initial δ function term needs to be added to represent the case when there is only one family. The second line is obtained merely by rearranging the terms in the sum. We have done this to demonstrate that this is the same as the result of [34] (their Eq. 34). In the above equation $\phi_k(x)$ and $Z_k(\tau)$ are given by

$$\phi_k(x)=(k-1)(1-x)^{k-2}\,, \tag{14a}$$

$$Z_k(\tau)=\frac{1}{k!(k-1)!}\sum_{i=k}^{\infty}(-1)^{i+k}(2i-1)\frac{(i+k-2)!}{(i-k)!}\times\exp[-i(i-1)\tau/2]\,. \tag{14b}$$

The latter result has also been derived using coalescent theory [22] and is equivalent to Eqs. (5.5) and (6.3) of Tavaré [23]. In this problem τ plays a similar role to θ in the infinite-alleles model, but while the number of τ families is typically fairly small if τ is of order 1, the number of alleles in the population in the infinite-alleles model diverges as $N\to\infty$ with θ constant. Thus there is a qualitative difference between the $f(x)$ distributions in (13) and (5).

The Y distribution is also of interest in the τ-family problem. The result of Ref. [34] may be written as

$$\Pi(Y)=Z_1(\tau)\delta(Y-1)+\sum_{k=2}^{\infty}Z_k(\tau)\Pi_k(Y)\,, \tag{15}$$

$$\Pi_k(Y)=(k-1)!\int_0^1 dx_1\cdots\int_0^1 dx_k\,\delta(1-x_1-\cdots-x_k)\times\delta(Y-x_1^2-\cdots-x_k^2)\,. \tag{16}$$

Note that, since the probabilities $Z_k(\tau)$ of there being k families are dependent on τ, $\Pi(Y)$ is also dependent on τ. However, the distributions of Y given that there are k families are simply a set of universal functions $\Pi_k(Y)$ independent of τ. It is therefore of interest to look at the shapes of the $\Pi_k(Y)$ in more detail. It is simple to show that $\Pi_2(Y)=(2Y-1)^{-1/2}$ for $Y>\frac{1}{2}$. Obtaining an analytical form for higher values of k from Eq. (16) becomes rapidly rather difficult. There is, however, a very simple numerical technique that allows us to obtain the shape of the $\Pi_k(Y)$ for all k. First generate k independent random variables ξ_i with an exponential probability distribution $\rho(\xi)=\exp(-\xi)$ and then obtain their sum S. A sample of family sizes with the required distribution can then be obtained by letting $x_i=\xi_i/S$. This can be checked by calculating the distribution of the x_i obtained by this procedure, in the following way. Let $z=\xi_2+\xi_3+\cdots+\xi_k$ and $S=\xi_1+z$. Since the ξ_i are independent, the distribution of z is $\rho_{k-1}(z)=e^{-z}z^{k-2}/(k-2)!$ and the distribution of x_1 is

$$\int d\xi_1\rho(\xi_1)\int dz\,\rho_{k-1}(z)\delta\left[x_1-\frac{\xi_1}{\xi_1+z}\right]=(k-1)(1-x_1)^{k-2}, \quad (17)$$

which is the required distribution $\phi_k(x_1)$ given in (14a). The appropriate joint distributions for two or more x_i generated by this method can also be calculated and shown to be correct. In Fig. 2 the $\Pi_k(Y)$ have been obtained by generating samples according to the above procedure. These distributions have singularities at $Y=\frac{1}{2},\frac{1}{3},\ldots$ as before.

Derrida has stressed [35] that many problems of this type can be reduced to a sum of independent random variables which is subsequently normalized, as was done above. For example, the mean field spin glass can be reduced to the random energy model, which in turn is equivalent to the sum of variables with a Levy distribution [36]. For the spin glass and for the infinite-alleles model there is an infinite number of variables, but only the few largest x_1 contribute to Y, since they are much larger than the typical x_i. In the τ-family problem this is not true. For a given k, all the k τ families are roughly the same size. The average of Y conditional on k can be shown to be $\int\Pi_k(Y)Y\,dY=2/(k+1)$, which goes to zero for large k.

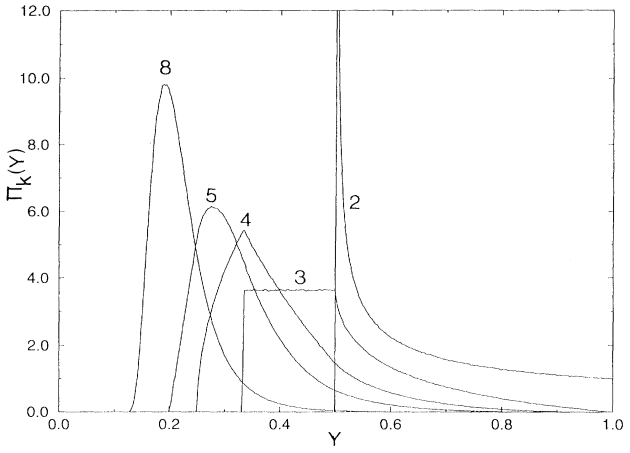

FIG. 2. A set of universal functions $\Pi_k(Y)$ arises in the τ-family problem on genealogical trees. These are calculated here for $k=2$, 3, 4, 5, and 8 using the method of summation of independent variables described in the text.

Since we now have the $\Pi_k(Y)$, it is possible to obtain the full Y distribution for a given τ using (15). Terms up to $k=12$ have been included in the curves of Fig. 3. These are compared to curves obtained from simulating the genealogy for small populations. For these simulations we used the matrix of times T_{ij} since each pair of individuals i and j had a common ancestor. At each subsequent generation this matrix is obtained from the preceding generation using the rules discussed in Refs. [8] and [32]. The matrix contains sufficient information to determine the τ-family sizes for any time τ. Since this procedure is relatively slow, one is forced to use a fairly small population size and one is therefore limited to a fairly small number of bins in the probability distribution for Y, as can be seen in Fig. 3. Agreement with the theory expected for infinite populations is nevertheless quite good.

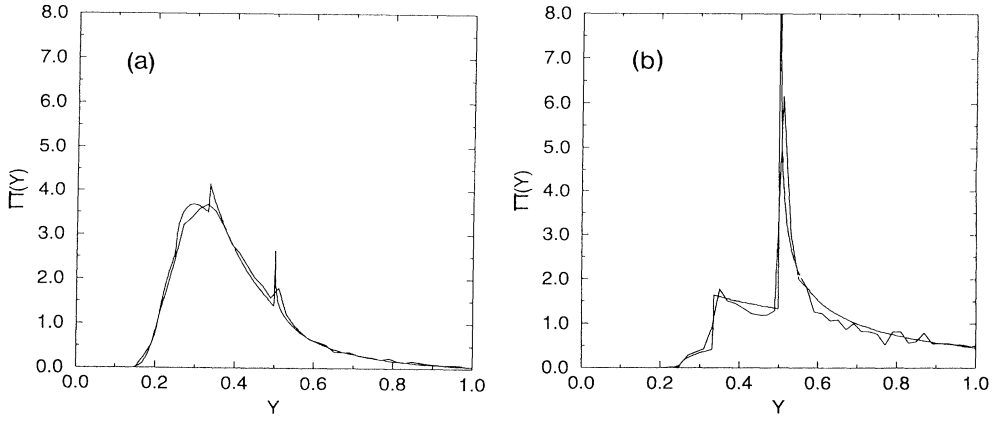

FIG. 3. $\Pi(Y)$ is shown for the τ-family problem for (a) $\tau=\frac{1}{2}$ and (b) $\tau=1$. The smooth curves with the sharp singularities are calculated using the functions $\Pi_k(Y)$ according to Eq. (15). The noisy curves are obtained from simulation of the genealogy and have considerable finite size effects.

IV. SAMPLING

We will now return to the infinite-alleles model and consider the problem of sampling. Suppose we wish to estimate the homozygosity Y in a population at a given locus. It is not possible to test all the individuals in the population. Typically we can test n individuals, where $N \gg n \gg 1$. Suppose that, in the n-alleles tested, there are k distinct allele types and n_i alleles of type i ($i=1,2,\ldots,k$). Let $Y^{(n)}$ be the estimate of Y obtained from the sample: $Y^{(n)}=\sum_{i=1}^{k}(n_i/n)^2$. As long as n is fairly large, $Y^{(n)}$ will be a good estimate of Y for the whole population. This is because Y is dominated by the high-frequency alleles and one can obtain a good estimate of the frequency of these high-frequency alleles by testing only a small fraction of the population. Many of the very low-frequency alleles in the population will not be present at all in the sample of n, but this does not matter since they do not contribute much to Y. The joint probability distribution of the set of n_i and k is known as the Ewens sampling formula [37,21,31]. Ewens has also calculated the probability that there are k distinct alleles in the sample. We will denote this probability $Z_k(\theta,n)$ since we wish to underline the similarity with the $Z_k(\tau)$ in the τ-family problem. From [37] and [21] (Sec. 3.6) we have

$$Z_k(\theta,n)=|S_n^k|\theta^k/S_n(\theta)\ , \tag{18}$$

where $S_n(\theta)=\theta(\theta+1)(\theta+2)\cdots(\theta+n-1)$ and $|S_n^k|$ is the coefficient of θ^k in $S_n(\theta)$. Note that the mutation rate and the overall population size enter into the problem only through the parameter θ. It turns out that although the distribution of the n_i depends on θ, the distribution of the n_i *conditional on there being k alleles in the sample* does not depend on θ. One consequence of this is that we may write

$$\Pi(Y^{(n)})=Z_1(\theta,n)\delta(Y^{(n)}-1)+\sum_{k=2}^{\infty}Z_k(\theta,n)\Pi_k(Y^{(n)})\ , \tag{19}$$

where the similarity to Eq. (15) is obvious. $\Pi(Y^{(n)})$ is the overall distribution of $Y^{(n)}$, which is dependent on θ, while $\Pi_k(Y^{(n)})$ is the distribution conditional on k, which does not depend on θ. This is important if one wishes to test the neutral theory. Watterson [29] calls θ a "nuisance parameter" since its value may not be known very well. It is possible to test whether data are consistent with the neutral theory using the functions Π_k without knowing θ. Simulated curves for the functions $\Pi_k(Y^{(n)})$ are given in Ref. [29].

Derrida and Flyvbjerg [17] finish their article by asking whether the singularities in the $\Pi(Y)$ distribution might be observable in any measurable quantity. Models such as the random map and the randomly broken object are too abstract to have a direct physical realization and most of the theory on spin glasses applies to mean field models that may be rather different from real three-dimensional spin-glass materials. The situation is different in biology, where gene frequencies and homozygosities are directly measurable. Comparisons between measured distributions of Y and the predictions of the neutral theory have been given by Fuerst, Chakraborty, and Nei [38] and Singh and Rhomberg [39]. The evidence that these studies provide as regards the validity of the neutral hypothesis has been reviewed by Kimura [40] and Gillespie [41]. In Fig. 4 we have plotted data taken from Fig. 5 of Ref. [39] for the homozygosity distribution of 61 polymorphic loci in *Drosophila melanogaster* (this data set is also reproduced by Gillespie [41]). The histogram has been reversed since it was plotted as heterozygosity $(1-Y)$ in the original paper. The vertical scale has been converted to probability density. The mean value of Y in the sample is close to 0.75, which from Eq. (8) corresponds to $\theta=\frac{1}{3}$. The theoretical curve is for the infinite-alleles model with the same value of θ, calculated by the iteration procedure described above. A quantitative estimation of the goodness-of-fit given in the Appendix indicates that the fit is moderately good. However, it is not our purpose here to argue for or against the neutral theory. We merely wish to stress that Y, beloved of theoretical physicists, is an experimental observable in biology. At the very least, we can say from Fig. 4 that the observed distribution is bimodal: there are many loci with Y close to 1 and many close to $\frac{1}{2}$. Fuerst, Chakraborty, and Nei [38] also comment on the peaks observed in their Y distributions. This may be considered to amount to an experimental observation of the singularities in $\Pi(Y)$ and we may expect this demonstration to become more convincing with the increasing availability of experimental data. Figure 4 also stresses the large variations in Y which are to be expected between different loci because the evolutionary process is non-self-averaging.

In summary, we have shown that the infinite-alleles model is identical to the randomly broken object problem, that the annealed random map is identical to the problem of population genealogies, and that there is a large number of other problems which are mathematically very similar. We find it fascinating that there should

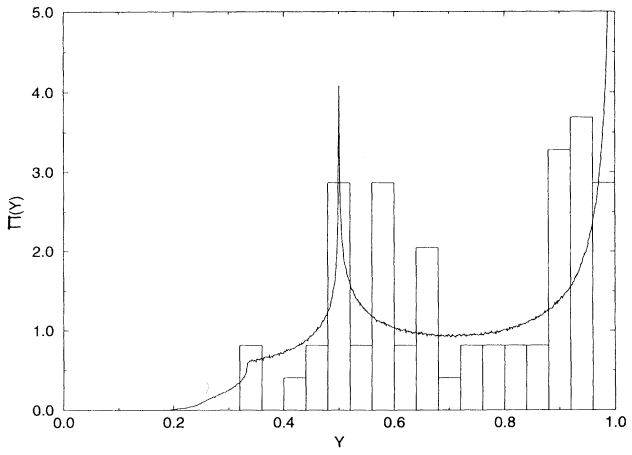

FIG. 4. Histogram of homozygosity for 61 loci in *Drosophila*, taken from Ref. [39], compared to the theoretical curve for the infinite alleles model. There is only one free parameter θ, which has been set to $\frac{1}{3}$, so that the mean value of Y for the theoretical curve is the same as that for the data.

be such a large number of close parallels between statistical physics and population genetics and we hope that this article goes some way towards bridging the gap.

APPENDIX: QUALITY OF FIT OF THEORY TO DATA

The experimental data in Fig. 4 have been divided into $B=25$ frequency boxes. The theoretical probability p_i of obtaining a Y value in the ith frequency box is known from the $\Pi(Y)$ curve. The probability of obtaining a sample with $m_1, m_2, \ldots, m_B$ loci falling in boxes $1,2,\ldots,B$ is

$$P(m_1,m_2,\ldots,m_B)=p_1^{m_1}p_2^{m_2}\cdots p_B^{m_B}\frac{M!}{m_1!m_2!\ldots m_B!}\ , \tag{A1}$$

where the total number of loci tested is $M=61$ in this example. The probability of obtaining the particular data set which was measured can be calculated to be $P_{\text{expt}}=1.1\times10^{-16}$. We generated random samples of 61 genes according to the theoretical distribution and calculated P for each one. A total of 100 000 samples were generated and 3% of these were found to have $P\leq P_{\text{expt}}$. This may be considered as, at best, only a moderately good fit of the data to the model.

We do not consider these data to be sufficient to make any strong claims with regard to the validity of the neutral theory. One problem with fitting the data in this way is that it is assumed that all loci have the same mutation rate and hence the same value of θ. This may well not be the case. A way of avoiding this problem is to use the Y distributions conditional on k, as was proposed by Watterson [29] and discussed in Sec. IV above. There are several other potential problems with the data: the population size may not have remained constant, the experimental technique may not detect all of the different alleles that are present, the choice of loci for analysis may have been biased, etc. (see [41]). In view of all this, Fig. 4 appears to be a surprisingly good fit. As stated previously, the main reason for showing this curve is to demonstrate to physicists that $\Pi(Y)$ is a measurable quantity, rather than to enter into the technicalities of data fitting.

[1] S. A. Kauffman, *The Origins of Order: Self-Organization and Selection in Evolution* (Oxford University Press, London, 1993).
[2] C. Amitrano, L. Peliti, and M. Saber, J. Mol. Evol. **29**, 513 (1989).
[3] W. Fontana, P. F. Stadler, E. G. Borberg-Bauer, T. Griesmacher, I. L. Hofacker, M. Tacker, P. Tarazona, E. D. Weinberger, and P. Schuster, Phys. Rev. E **47**, 2083 (1993).
[4] P. G. Higgs, J. Phys. I (France) **3**, 43 (1993).
[5] R. Rammal, G. Toulouse, and M. A. Virasoro, Rev. Mod. Phys. **58**, 765 (1986).
[6] B. Derrida and L. Peliti, Bull. Math. Biol. **53**, 355 (1991).
[7] P. G. Higgs and B. Derida, J. Phys. A **24**, L985 (1991).
[8] P. G. Higgs and B. Derrida, J. Mol. Evol. **35**, 454 (1992).
[9] P. Bak and K. Sneppen, Phys. Rev. Lett. **71**, 4083 (1993).
[10] H. Flyvbjerg, K. Sneppen, and P. Bak, Phys. Rev. Lett. **71**, 4087 (1993).
[11] M. Mézard, G. Parisi, N. Sourlas, G. Toulouse, and M. Virasoro, J. Phys. (Paris) **45**, 843 (1984).
[12] M. Mézard, G. Parisi, and M. A. Viarasoro, J. Phys. (Paris) Lett. **46**, L217 (1985).
[13] B. Derrida and G. Toulouse, J. Phys. (Paris) Lett. **46**, L223 (1985).
[14] S. A. Kauffman, J. Theor. Biol. **22**, 437 (1969).
[15] H. Gutfreund, J. D. Reger, and A. P. Young, J. Phys. A **21**, 2775 (1988).
[16] B. Derrida and H. Flyvbjerg, J. Phys. (Paris) **48**, 971 (1987).
[17] B. Derrida and H. Flyvbjerg, J. Phys. A **20**, 5273 (1987).
[18] S. Wright, *Evolution and the Genetics of Populations Vol. 2: The Theory of Gene Frequencies* (University of Chicago Press, Chicago, 1969).
[19] M. Kimura, J. Appl. Prob. **1**, 177 (1964).
[20] J. F. Crow and M. Kimura, *An Introduction to Population Genetics Theory* (Harper & Row, New York, 1970).
[21] W. J. Ewens, *Mathematical Population Genetics* (Springer-Verlag, Berlin, 1979).
[22] J. F. C. Kingman, J. Appl. Prob. **19A**, 27 (1982).
[23] S. Tavaré, Theor. Pop. Biol. **26**, 119 (1984).
[24] P. Donnelly and S. Tavaré, Adv. Appl. Prob. **18**, 1 (1986).
[25] P. Joyce and S. Tavaré, Stoch. Proc. Appl. **36**, 245 (1990).
[26] F. M. Stewart, Theor. Pop. Biol. **9**, 188 (1976).
[27] G. A. Watterson, Genetics **88**, 405 (1978).
[28] R. H. MacArthur, Proc. Natl. Acad. Sci. U.S.A. **43**, 293 (1957).
[29] G. A. Watterson, Theor. Pop. Biol. **6**, 217 (1974).
[30] G. P. Patil and C. Taillie, Bull. Int. Stat. Inst. **47** (2), 497 (1977).
[31] P. Donnelly, Theor. Pop. Biol. **30**, 271 (1986).
[32] P. G. Higgs and G. Woodcock, J. Math. Biol. (to be published).
[33] S. Tavaré, W. J. Ewens, and P. Joyce, Genetics **122**, 705 (1989).
[34] B. Derrida and D. Bessis, J. Phys. A **21**, L509 (1988).
[35] B. Derrida (unpublished).
[36] J. P. Bouchaud and A. Georges, Phys. Rep. **195**, 127 (1990).
[37] W. J. Ewens, Theor. Pop. Biol. **3**, 87 (1972).
[38] P. A. Fuerst, R. Chakraborty, and M. Nei, Genetics **86**, 455 (1977).
[39] R. S. Singh and L. R. Rhomberg, Genetics **117**, 255 (1987).
[40] M. Kimura, *The Neutral Theory of Molecular Evolution* (Cambridge University Press, Cambridge, England, 1983).
[41] J. H. Gillespie, *The Causes of Molecular Evolution* (Oxford University Press, London, 1991).