Cluster distributions in physics and genetic diversity

A. Z. Mekjian
Institute for Nuclear Theory, HN-12, University of Washington, Seattle, Washington 98195
and Department of Physics and Astronomy, Rutgers University, Piscataway, New Jersey 08854
(Received 5 February 1991)

An underlying similarity in the mathematical structure between a problem in the physical sciences and a problem in the biological sciences is presented. An isomorphism is established between cluster distributions in physics and issues related to genetic diversity in biology as given by the Ewens sampling theory [Theor. Population Bio. **3**, 87 (1972)]. Allelic or genetic diversity, as measured by the number and frequency of different alleles (gene types), has a correspondence with the size distribution of clusters in physics. The rate of mutation in genetics is shown to have its parallel in Richardson's thermionic emission rate in physics. Using methods from combinatorial analysis and from the symmetric group S_n, simple formal connections between these two areas are developed. The logarithmic series of Fisher, Corbet, and Williams [J. Animal Ecol. **12**, 42 (1943)] for species abundance appears in the solution developed as does a scale-invariant hyperbolic function. Maximum entropy methods are discussed. Even though the underlying dynamical processes are quite different in these two areas, new insights and the possibility that methods developed in one area may be used with advantage in the other area may follow from this correspondence.

PACS number(s): 87.10.+e, 25.70.Np, 05.20.Dd

I. INTRODUCTION

In this paper a formal correspondence is established between a cluster distribution in the physical sciences and an allelic distribution in genetics. A cluster-size distribution function is a relationship between the number of clusters of a given size versus the size of the cluster. For example, a collision between two nuclei produces fragments of varying sizes. The size of a cluster is given by its mass number or the number of nucleons contained in it (number of protons plus neutrons). The number of fragments or nuclei n_i with a given mass number i produced in such a fragmentation process is a distribution function for cluster sizes. As an illustration, three carbon nuclei may be present, each containing 12 nucleons or six protons plus six neutrons.

Genetic diversity in the biological sciences is measured by the number of different alleles (i.e., types of genes or DNA sequences) which occur at a given gene locus, together with the frequencies of these alleles. This diversity can usually only be estimated by taking a sample of genes from the population, and is described by the number a_i of alleles in the sample which are represented in that sample by exactly i genes. For example, if $a_3=6$, there are six different alleles (gene types) each appearing exactly three times in the sample.

The correspondence that is established in Sec. II is between the distribution of cluster sizes n_i and the genetic diversity distribution a_i as given by the Ewens sampling theory [1]. Specifically, a one-to-one correspondence is established between $n_i$clusters of size i and a_i different alleles each appearing i times. While n_i is a measure of cluster similarity in a cluster distribution (how many fragments are the same by having the same atomic number or mass), the a_i is a measure of genetic diversity (how many different alleles each appear i times). The transcription just given relates similarity in the domain of the physical sciences to diversity in the realm of the biological sciences. A quantity that measures allelic similarity, called the homozygosity, will be shown to be related to moments of the cluster-size distribution function. However, it should be emphasized that the formal correspondence between these two areas does not imply some underlying similarities in the basic processes. What may be achieved in the formal correspondence are new insights in one area derived from the other area and the possibility that methods developed in one area may also be used with advantage in the other area.

The interrelationship discussed above arose out of a recent realization that a particular simple model of fragmentation developed in Refs. [2,3] contained a logarithmic series for the distribution of clusters. A logarithmic series was introduced by Fisher [4] in a study of the distribution of butterflies and moths caught by light traps. In particular, the number of different types or species of butterflies and moths observed when plotted as a function of frequency of appearance in a sample could be described as a logarithmic series. In fact the species diversity versus frequency of occurrence was very close to being a hyperbolic function. The model in Refs. [2,3] contained an exact hyperbolic behavior at a particular point. Further investigations showed that the recent simple model of fragmentation had a formal mathematical structure very similar to a much earlier genetic sampling theory pioneered by Ewens [1]. The Ewens theory has been developed over the years since its introduction—see Refs. [5–17] and references in these papers. It should be noted that questions of genetic diversity and the Ewens sampling formulas are not directly connected with species diversity.

From the observation of this similarity in mathemati-

cal structure of these two approaches a direct correspondence can be made between the two areas. This paper explores some of the results of this direct correspondence between the Ewens theory and cluster theories. Furthermore, using this correspondence more complex theories [18–24] of cluster distributions may be useful in an investigation of genetic diversity. The application of more complex theories will be developed to some extent in this paper.

An outline of this paper is as follows. Section II discusses an interrelationship between cluster distributions and genetic diversity. In Sec. II A the partitioning of objects into clusters of various sizes is compared to the grouping of alleles into classes according to frequency of occurrence. Then, in Secs. II B nd II C, the cluster model in Refs. [2,3] is compared to the Ewens model [1]. A general approach to either problem is developed in Sec. II D in terms of a generating function or, equivalently, a grand canonical ensemble. Next, in Sec. II E, a simple solution for the mean number of clusters of a given size found in Refs. [2,3] is applied to various issues related to genetic diversity. The Fisher logarithmic series is shown to be contained in the simple, exact solution as an approximation. A maximum entropy approach (Sec. II G) is also developed and shown to give a solution which is close to the exact solution. Using the formal correspondence between the two areas various concepts that appear in cluster theories are used to further investigate genetic diversity. These include simple and factorial moments of the distribution which are discussed in Secs. II F and II H. Solutions for more complex weight factors are given in Sec. II I and equilibrium and nonequilibrium distributions are discussed in Sec. II J.

Section III connects a simple, exact solution for allelic diversity or cluster distributions to expressions found in probability theory. Specifically, these distributions have simple interpretations in terms of Bernoulli trials and in terms of urn models with replacement. Section IV summarizes this paper.

II. CLUSTER DISTRIBUTION AND GENETIC DIVERSITY

A. Partitions

Both cluster and allelic distributions consider the partitioning of a fixed number of objects into groups. For the cluster situation, A objects (nucleons) are partitioned into groups specified by the cluster size k and the number of clusters of that size, called n_k. A constraint $A=\sum_k k n_k$ is imposed. In the allele case, the partitioning is by the number of times any allelic type occurs. As noted above, a quantity a_i is defined to be the number of allelic types represented by i genes in a sample of n genes. Then $n=\sum_i i a_i$. Both the n_k's and a_i's form a partition:

$$\mathbf{n}=(n_1,n_2,\ldots,n_k,\ldots,n_A)$$
$$=(1^{n_1},2^{n_2},\ldots,k^{n_k},\ldots,A^{n_A}) \quad (2.1)$$

or

$$\mathbf{a}=(a_1,a_2,\ldots,a_i,\ldots,a_n)$$
$$=(1^{a_1},2^{a_2},\ldots,i^{a_i},\ldots,n^{a_n}). \quad (2.2)$$

Figure 1(a) illustrates a specific case. The A or n can be thought of as blocks and a partition can be represented as a block diagram as shown in Fig. 1(b). Such partitions appear in number theory and represent the number of ways a given integer N can be decomposed into integer summands such as $5=4+1=3+2=3+1+1=2+2+1=2+1+1+1=1+1+1+1+1$. In the cluster case the numeral 2 will represent a cluster of size 2, while in the genetic case the numeral 2 will represent a given gene which appears twice. The partition $5=2+2+1$ corresponds to two different alleles each appearing twice and a single allele, or two clusters of size 2 and one monomer. The multiplicity

$$m=\sum_k n_i=\sum_i a_i \quad (2.3)$$

is either the total number of fragments (clusters) or the total number of different varieties of alleles.

A plot of a_i versus i is then a representation of the number of different allelic types observed in the sample as a function of how often each occurs from singletons to the most frequently occurring ones. The most frequent alleles are those with many copies of themselves. A plot of n_k versus k is a representation of the number of clusters of a given size versus their size. The total number of partitions is just the total number of ways of arranging N ($N=A=n$) blocks into columns and rows, one of which is shown in Fig. 1(b).

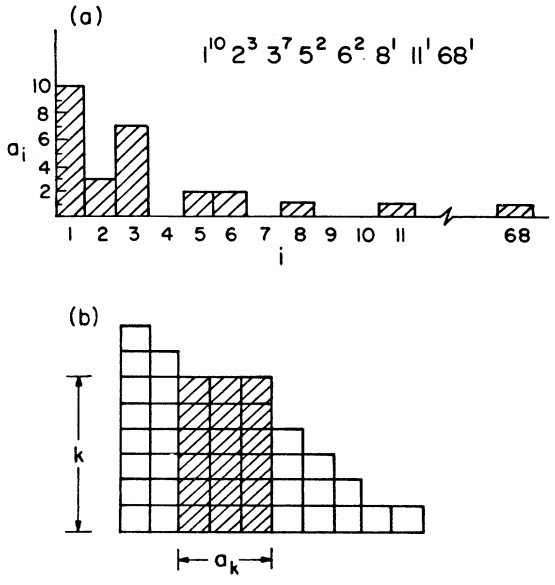

FIG. 1. Partitions and block diagrams. (a) illustrates an allelic partition $1^{10},2^3,3^7,5^2,6^2,8^1,11^1,68^1$. The partition is from Singh, Lewontin, and Felton (Ref. [25]). This partition represents ten singleton genes, three different alleles each appearing twice, seven different alleles represented three times , . . . , and one allele represented 68 times. (b) is a block diagram for a partition. The vertical axis is k and the horizontal axis is n_k or a_k.

The total number of possible arrangements $P(n)$ can be obtained from a power series generating function [25]

$$\frac{1}{(1-x)(1-x^2)\cdots(1-x^k)\cdots}=\sum_{n=1}^{\infty}p(n)x^n \ . \tag{2.4}$$

The factor $(1-x^k)^{-1}=1+x^k+x^{2k}+\cdots$ counts objects of size k appearing once, x^k, twice, x^{2k}, etc. The coefficient of x^n, the $p(n)$ in Eq. (2.4), gives all the possible ways of obtaining n in terms of integer summands. An example is the case $n=5$ given above which has $p(5)=7$.

The decomposition of n into integer summands with the total number of parts fixed, called m in Eq. (2.3), is $p(n,m)$. The $p(n,m)$ satisfies a recurrence relationship [26] $p(n,m)=p(n-1,m-1)+p(n-m,m)$ and $p(n)=\sum_{m=1}^{n}p(n,m)$. When n is large, asymptotic expressions for $p(n)$ and $p(n,m)$ are useful. The Hardy-Ramanujan formulas [26] are for $p(n)$,

$$p(n)\sim\frac{e^{\pi\sqrt{2n/3}}}{n4\pi\sqrt{3}} \ , \tag{2.5}$$

and for $p(n,m)$, for n large and m not too large $[m\sim O(n^{1/3})]$,

$$p(n,m)\sim\frac{n^{m-1}}{m!(m-1)!} \ . \tag{2.6}$$

Both $p(n)$ and $p(n,m)$ increase rapidly with increasing n.

B. Partition weights

For each partition discussed in Sec. II A, a weight is assigned. The Ewens theory [1] has a probability assignment for each partition $\mathbf{a}=(a_1,a_2,\ldots,a_n)$ given by

$$p(\mathbf{a},\theta)=\left[\frac{n!}{a_1!1^{a_1}a_2!2^{a_2},\ldots,a_n!n^{a_n}}\right]\frac{\theta^m}{L(\theta)} \ . \tag{2.7}$$

The $L(\theta)=\theta(\theta+1)\cdots(\theta+n-1)$.

The Ewens distribution of Eq. (2.7) is based on the infinitely many alleles model, which arises from the recognition of the gene as a DNA sequence admitting an extremely large number of different sequence possibilities. In the infinitely many alleles model every mutation produces a new allelic type or DNA sequence not already present in the existing population. The distribution of Eq. (2.7) arises, at a steady state, between the force of mutation (creating and increasing genetic diversity) and the action of random genetic drift (eliminating variations). The model assumes non-Darwinian behavior, i.e., that the different allelic types are selectively equivalent. The parameter θ was shown [1] to be defined from the effective population size N_e and the mutation rate u (both unknown) through the formula:

$$\theta=4N_eu \ . \tag{2.8}$$

Details of the Ewens model can be found in Refs. [6–17,27].

In one of the fragmentation models considered [2,3], a similar probability assignment for a partition into clusters $\mathbf{n}=(n_1,n_2,\ldots,n_A)$ was used:

$$p(\mathbf{n},x)=\frac{A!}{n_1!1^{n_1}n_2!2^{n_2}\cdots n_k!k^{n_k}\cdots}\frac{x^m}{L(x)} \tag{2.9}$$

with $L(x)=x(x+1)\cdots(x+A-1)$. The parameter x was shown to be [2,3]

$$x=\frac{V}{v_0}e^{-a_B/k_BT}e^{-k_BTT_0/(T+T_0)\varepsilon_0} \ . \tag{2.10}$$

Here V is the volume of the system in which a quasiequilibrium is established and v_0 is a quantum volume associated with a thermal wavelength λ_T^3,

$$v_0=\lambda_T^3=\frac{h^3}{(2\pi mk_BT)^{3/2}} \ , \tag{2.11}$$

where h is Planck's constant, m the mass of a nucleon or monomer, and k_B is Boltzmann's constant. The a_B is a coefficient which appears in the binding energy $E_B(k)$ of a cluster of size k. Namely, $E_B(k)=a_B(k-1)$. The ε_0 is the level spacing in the density of excited states in a cluster of size k. The T_0 is a cutoff temperature for internal excitations. Further details can be found in Refs. [2,3].

The probability function of Eq. (2.9) was arrived at using entropy arguments, where the entropy associated with a given partition $\mathbf{n}$ was developed based on the Sakur-Tetrode law of thermodynamics [28]. The factorials $n_k!$ which appear in Eq. (2.9) are Gibbs factorials. These factorials arose in a prescription developed by Gibbs to remove problems associated with the entropy of mixing [28]. Terms such as k^{n_k} which appear in Eq. (2.9) arose from internal excitations and such terms are discussed in more detail in the Appendix.

Once an entropy assignment $S(\mathbf{n},x)$ is given to a partition $\mathbf{n}$, a weight factor $W(n,x)$ can be assigned to that partition:

$$W(\mathbf{n},x)=e^{S(\mathbf{n},x)/k_B} \ . \tag{2.12}$$

A quantity $\sigma(\mathbf{n},x)\equiv S(\mathbf{n},x)/k_B$ can be introduced which is a dimensionless entropy in the cluster case. A normalized probability assignment is then

$$p(\mathbf{n},x)=\frac{W(\mathbf{n},x)}{D(x)} \ , \tag{2.13}$$

where

$$D(x)=\sum_{\pi(\mathbf{n})}W(\mathbf{n},x) \ . \tag{2.14}$$

The sum is over all partitions $\mathbf{n}$ of $A=\sum_k kn_k$ which is called $\pi(n)$. There are $p(n)$ terms in the sum. Given Ewens's probability assignment, a dimensionless functional $\sigma(\mathbf{a},\theta)$ for a given partition $\mathbf{a}$ can be obtained:

$$\sigma(\mathbf{a},\theta)=f(n)\ln\left[\prod_i\frac{\theta^m}{a_i!i^{a_i}}\right] \ . \tag{2.15}$$

The $f(n)$ is an arbitrary function of $n=\sum_i ia_i$ and is partition independent. A factor $D(\theta)=\theta(\theta+1)\cdots(\theta+n-1)\exp f(n)$ normalizes the weight $W(\mathbf{a},\theta)=\exp\sigma(\mathbf{a},\theta)$ to a probability function:

$$p(\mathbf{a},\theta)=\frac{e^{\sigma(\mathbf{a},\theta)}}{D(\theta)} . \tag{2.16}$$

The functionals $\sigma(\mathbf{a},\theta)$ and $S(\mathbf{n},x)$ will be used later and these results will be generalized. A class of different probability assignments will be shown to be calculable and also maximum entropy principles will be used to obtain a solution for generalized weight assignments.

This section ends with the formal correspondence obtained by comparing Eqs. (2.7) and (2.9):

$$\begin{aligned} &A \leftrightarrow n , \\ &n_k \leftrightarrow a_k , \\ &x \leftrightarrow \theta . \end{aligned} \tag{2.17}$$

Later, it will be shown that this equivalence can be interpreted as an isomorphism within the framework of the symmetric group S_n.

C. Correspondence between x and θ

The fragmentation parameter x is given by Eq. (2.10) and this parameter will now be recast into a form that involves an evaporation rate. The diversity parameter θ of Eq. (2.8) involves a mutation rate u. To rewrite x into a form involving a rate, a characteristic speed v_s can be introduced which is associated with the temperature T and is $v_s=\sqrt{\alpha kT/m}$. When v_s is taken as the root-mean-square speed of a Maxwell-Boltzmann velocity distribution, then $\alpha=3$ since $\langle mv^2/2\rangle=3k_BT/2$ by equipartition. For the most probable speed $\alpha=2$, and for the mean speed $\alpha=8/\pi$. Taking the volume $V=SD/6$, where S is the surface area associated with V and D its diameter, a characteristic transit time across a distance D is $t=D/v_s$. Then x is

$$x=tS[Je^{-k_BTT_0/(T+T_0)\varepsilon_0}] . \tag{2.18}$$

A factor $\sqrt{2\pi\alpha}/6\sim 1$ has been omitted. The quantity J,

$$J=\frac{2\pi m(k_BT)^2}{h^3}e^{-a_B/k_BT} , \tag{2.19}$$

which appears in the square brackets of Eq. (2.18), is Richardson's formulas [29] (excluding a spin factor) for the evaporation rate per unit area of particles from a heated metal. The work function W_f which normally appears in the Richardson formula is here a_B. The work function or separation energy for a binding energy expression $E_B(k)=a_B(k-1)$ turns out to be just a_B. At low T, the quantity $\exp[-k_BTT_0/(T+T_0)\varepsilon_0]$ can be neglected compared to $\exp(-a_B/k_BT)$. For high T, the metal vaporizes.

D. Isomorphisms and generalized weight factors based on the permutation or symmetric group S_n

The combinatorial prefactor which appears in both Eqs. (2.7) and (2.9) appears in permutation problems [30] and in the group structure of S_n [31]. This observation will be used first to establish a simple isomorphism between genetic diversity and cluster distributions, and second to generalize the weight factor of Sec. II B. The connection between the combinatorial prefactor and permutation problems was also noted in Watterson [7] and Kingman [32] for the Ewens theory. References [2,3] used this connection to obtain simple expressions for the cluster-size distribution function. Also, Ref. [24] uses the connection with S_n to obtain general expressions for the distribution of cluster sizes.

For n objects, $n!$ possible permutations are possible. The set of all permutations forms a non-Abelian group S_n. The group S_n can be divided into conjugacy classes according to cycle structure. These classes are specified by the length of a cycle j and the number of such cycles. For example, the permutation

$$\begin{bmatrix}1&2&3&4&5\\2&1&3&5&4\end{bmatrix}=\begin{bmatrix}1&2\\2&1\end{bmatrix}\begin{bmatrix}3\\3\end{bmatrix}\begin{bmatrix}4&5\\5&4\end{bmatrix} \tag{2.20}$$

has two cycles of length 2 and one cycle of length 1. In general there will be n_1 cycles of length 1, n_2 cycles of length 2, . . . , n_j cycles of length j, ... A quantity

$$M_2=\frac{n!}{n_1!1^{n_1}n_2!2^{n_2}\cdots n_j!j^{n_j}} , \tag{2.21}$$

which is called a Cauchy number, counts how many group elements have this cycle class structure. Since every permutation belongs to one and only one cycle class, the sum of the Cauchy number M_2 over all partitions is $n!$. The block diagram of Fig. 1 when rotated appears in Young tableaus for S_n. The sum of M_2 over all partitions with fixed $m=n_1+n_2+\cdots+n_n$ is a Stirling number S_n^m of the first kind, $(-1)^{n-m}S_n^m=\sum_{\pi(n,m)}M_2$, where the sum is over all n_i's constrained by both n and m. The $p(n)$ of Eq. (2.4) counts the number of cycle classes for the group S_n.

Since M_2 appears in the weight factor of Eqs. (2.7) and (2.9) a simple isomorphism can be established between cycle classes, allelic variations, and cluster distributions. For example, n_j cycles of length j are equivalent to a_j different alleles, each appearing j times, or n_j clusters of size j. The M_2 of Eq. (2.21) are combinatoric factors underlying the division into clusters, alleles, or cycle classes. In the group S_n, M_2 permutations have the cycle structure given by the partition formed by the n_j's. Figure 2 illustrates the correspondence stated for the case $n=5$. Permutation distributions are given by Eq. (2.7) for the particular choice $\theta=1$.

General weight functions, which contain M_2 as their combinatoric factor, can be developed as follows. Since there are n different length cycles ($j=1,2,\ldots,n$), a variable x_j can be assigned to tag each cycle. Let $\mathbf{x}=(x_1,x_2,\ldots,x_n)$, and for a particular partition $\mathbf{a}=(a_1,a_2,\ldots,a_n)$, in the a notation above, define a function

$$W_n(\mathbf{x},\mathbf{a})=\frac{n!}{\prod_j a_j!j^{a_j}}x_1^{a_1}x_2^{a_2}\cdots x_j^{a_j} . \tag{2.22}$$

The unnormalized weight $W_n(\mathbf{x},\mathbf{a})$ involves n variables, the x_j's which tag the length of the cycle or size of the

cluster or number of times an allele occurs.

Next consider the function

$$Q_n(\mathbf{x})=\sum_{\pi(n)} W_n(\mathbf{x},\mathbf{a}) , \tag{2.23}$$

where the sum is over all partitions of n, called $\pi(n)$. For example, when $x_1=x_2=\cdots=x_n=x$ and $\mathbf{x}\Longrightarrow x$, then

$$Q_n(x)=x(x+1)\cdots(x+n-1)=\frac{\Gamma(x+n)}{\Gamma(x)} , \tag{2.24}$$

where $\Gamma(x+n)$ is a gamma function; $\Gamma(m)=(m-1)!$ when m is an integer. Let

$$Q_n^m(\mathbf{x})=\sum_{\pi(n,m)} W_A(\mathbf{x},\mathbf{a}) , \tag{2.25}$$

where the sum is over all partitions of n with $m=\sum_i a_i$ also fixed. When all x_j's $= x$

$$Q_n^m(x)=(-1)^{n-m}S_n^m x^m . \tag{2.26}$$

The generating function for $Q_n(\mathbf{x})$ is [2]

$$\mathcal{L}(u,\mathbf{x})\equiv e^{ux_1+u^2(x_2/2)+u^3(x_3/3)+\cdots} =\sum_n Q_n(\mathbf{x})\frac{u^n}{n!} . \tag{2.27}$$

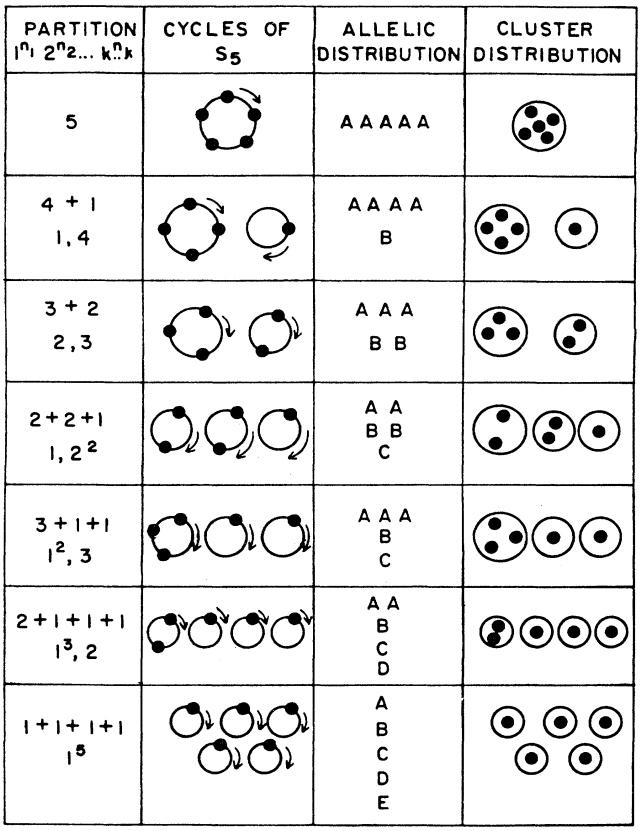

FIG. 2. Correspondence between cycles, alleles, clusters, and the decomposition of an integer. The case shown is for $n=5$. A cycle of length k is represented by k dots on a circumference. Different alleles are given the symbols A, B, C, D, and E. An allele which occurs twice is represented by AA, etc. A cluster of size k has k dots inside a circle.

Again when all the x_j's $= x$ then the left-hand side (lhs) is simply $1/(1-u)^x$ and $Q_n(x)$ is determined by expanding this expression and equating similar powers of u. The result is Eq. (2.24).

The $Q_n(\mathbf{x})$ can be considered a canonical partition function of statistical mechanics while

$$\mathcal{L}(u,\mathbf{x})=\sum_n Q_n(\mathbf{x})u^n/n!$$

is its corresponding grand canonical partition function. The $Q_n^m(\mathbf{x})$ of Eq. (2.25) is a more restricted partition function. Writing $u=e^{\beta\mu}$ where μ is a chemical potential, each canonical partition at fixed n, the $Q_n(\mathbf{x})$, is given a Boltzmann weight $e^{n\beta\mu}$. The $n!$ in Eq. (2.27) is a combinatoric factor. The quantities in $\mathcal{L}$ which have the factor $u^j=e^{\beta j\mu}=e^{\beta\mu_j}$, where $\mu_j=j\mu$, contain the chemical potentials μ_j for each species j. The condition $\mu_j=j\mu$ is a chemical equilibrium constraint. This last point will be returned to in Sec. II J.

Once a partition function is established, ensemble-averaged quantities follow by differentiating. Here, the results obtained in Refs. [23,24] will be listed using the transcription of notation given above. The mean number of alleles or clusters in the canonical ensemble (fixed n) is

$$\langle a_j\rangle=\sum_{\pi(a)} a_j P_n(\mathbf{a},\mathbf{x}) \tag{2.28}$$

which can be shown to be

$$\langle a_j\rangle=\frac{x_j}{j}\frac{n!}{(n-j)!}\frac{Q_{n-j}(\mathbf{x})}{Q_n(\mathbf{x})} . \tag{2.29}$$

The correlation $\langle a_i a_j\rangle$ reduces to

$$\langle a_i a_j\rangle=\frac{x_i x_j}{ij}\frac{n!}{(n-i-j)!}\frac{Q_{n-i-j}(\mathbf{x})}{Q_n(\mathbf{x})} . \tag{2.30}$$

The fluctuation $\langle a_j^2\rangle-\langle a_j\rangle^2$ can be obtained from

$$\langle a_j(a_j-1)\rangle=\frac{x_j^2}{j^2}\frac{n!}{(n-2j)!}\frac{Q_{n-2j}(x)}{Q_n(x)} \tag{2.31}$$

and the $\langle a_j\rangle$. Note the values of $i+j\le n$ and $2i\le n$ in $\langle a_i a_j\rangle$ and $\langle a_i^2\rangle$, otherwise the expectation values are zero since the sample size is exceeded. The results can easily be generalized to higher correlation functions. Similar results also hold for restricted partitions as in $Q_n^m(\mathbf{x})$:

$$\langle a_j\rangle_{\pi(n,m)}=\frac{x_j}{j}\frac{n!}{(n-j)!}\frac{Q_{n-j}^{m-1}(\mathbf{x})}{Q_n^m(\mathbf{x})} \tag{2.32}$$

for the fluctuations

$$\langle a_j(a_{j-1})\rangle_{\pi(n,m)}=\left[\frac{x_j}{j}\right]^2\frac{n!}{(n-2j)!}\frac{Q_{n-2j}^{m-2}(\mathbf{x})}{Q_n^m(\mathbf{x})} \tag{2.33}$$

and for the correlations

$$\langle a_i a_j\rangle_{\pi(n,m)}=\frac{x_i}{i}\frac{x_j}{j}\frac{n!}{(n-i)!(n-j)!}\frac{Q_{n-i-j}^{m-2}(\mathbf{x})}{Q_n^m(\mathbf{x})} , \tag{2.34}$$

where the above averages are over all partitions of $n=\sum_i i a_i$ with fixed $m=\sum_i a_i$, which are symbolically identified by the notation $\pi(n,m)$. The results of Eqs. (2.33) and (2.34) are also easily generalized.

E. Simplified model

This section gives an application of the results of the preceding section for a simple case, namely, when all the x_i's or θ_i's are equal. Only one variable will be used which is called θ. Then $Q_n(\mathbf{x})=Q_n(\theta)$ is simply $Q_n(\theta)=L(\theta)=\theta(\theta+1)\cdots(\theta+n-1)$ and $Q_n^m(\mathbf{x})=Q_n^m(\theta)=(-1)^{n-m}S_n^m\theta^m$ as already noted. All ensemble-averaged quantities then follow straightforwardly using the results of Eq. (2.29) through Eq. (2.34). For example, the mean number $\langle a_i\rangle$ has a very simple form:

$$\langle a_i\rangle=\frac{\theta}{i}\frac{n!}{(n-i)!}\frac{\Gamma(\theta+n-i)}{\Gamma(\theta+n)} . \tag{2.35}$$

The expectation of the product $a_i a_j$ is

$$\langle a_i a_j\rangle=\frac{\theta}{i}\frac{\theta}{j}\frac{n!}{(n-i-j)!}\frac{\Gamma(\theta+n-i-j)}{\Gamma(\theta+n)} \tag{2.36}$$

and correlations can be obtained from $\langle a_i a_j\rangle-\langle a_i\rangle\langle a_j\rangle\equiv C(a_i,a_j)$. The factorial product $a_i(a_i-1)$ has an expectation

$$\langle a_i(a_i-1)\rangle=\left[\frac{\theta}{i}\right]^2\frac{n!}{(n-2i)!}\frac{\Gamma(\theta+n-2i)}{\Gamma(\theta+n)} \tag{2.37}$$

and this result, along with Eq. (2.35), can be used to obtain the fluctuation $\langle a_i^2\rangle-\langle a_i\rangle^2$. The same procedure can be repeated for the restricted partition averages, specifically

$$\langle a_i\rangle_{\pi(n,m)}=\frac{1}{i}\frac{n!}{(n-i)!}\frac{|S_{n-i}^{m-1}|}{|S_n^m|} , \tag{2.38}$$

a result first obtained by Ewens [1] and Watterson [8]. The S_n^m and S_{n-i}^{m-1} are Stirling numbers introduced above [after Eq. (2.21)]. Moreover, using the procedure of the preceding section, fluctuations and correlations are also easily obtained from

$$\langle a_i(a_i-1)\rangle_{\pi(n,m)}=\frac{1}{i^2}\frac{n!}{(n-i)!}\frac{|S_{n-2i}^{m-2}|}{|S_n^m|} \tag{2.39}$$

and

$$\langle a_i a_j\rangle_{\pi(n,m)}=\frac{1}{i}\frac{1}{j}\frac{n!}{(n-i-j)!}\frac{|S_{n-i-j}^{m-2}|}{|S_n^m|} . \tag{2.40}$$

The $\langle a_i\rangle$ distribution of Eq. (2.35) contains a hyperbolic behavior at $\theta=1$, specifically $\langle a_i\rangle=1/i$ at $\theta=1$. The behavior of $\langle a_i\rangle$ with i is also similar to Fisher's logarithmic expression for species numbers as discussed below. Equation (2.38) has no hyperbolic point. A discussion of Fisher's distribution can be found in Watterson [8]. Again, it is emphasized that $\langle a_i\rangle$ refers to genetic or allelic diversity while Fisher's logarithmic series was developed for species diversity. Genetic diversity is not directly connected with species diversity.

Detailed properties of the behavior of $\langle a_i\rangle$ with θ can be found in Refs. [2,3]. Here, a brief summary will be stated which will be used later.

1. Solution at $\theta=0$

At $\theta=0$, the mutation rate or temperature ($\theta\leftrightarrow x$) is zero. Only one type of gene is present and it appears n times or has $n-1$ copies of itself. Also one cluster exists of size $k=n$.

2. Region $\theta\ll 1$

At low mutation rates or low temperatures T

$$\langle a_i\rangle=\frac{\theta n}{i(n-i)} \tag{2.41}$$

for $i=1$ to $n-1$, and $\langle a_n\rangle=1-\theta(\gamma+\ln n)$ where $\gamma=0.577\,22$ is Euler's constant. The diversity or cluster distribution is a U-shaped distribution in i, centered around $i\sim n/2$, except for the point $i=n$. As θ is increased $\langle a_n\rangle$ decreases and the back part of the U shape, representing frequently occurring alleles, or large clusters, also decreases. A good approximation to the behavior of $\langle a_i\rangle$ for $0<\theta<1$ is simply

$$\langle a_i\rangle=\frac{\theta}{i(1-i/n)^{1-\theta}} . \tag{2.42}$$

When the result of Eq. (2.10) for $x=\theta$ is substituted into Eq. (2.41), the resulting equation for $\langle a_1\rangle$, with $n\gg i=1$, is Fermi's [33] result for the evaporation of a particle into a cavity of volume V from a heated metal.

3. Power-law behavior at $\theta=1$

At $\theta=1$, the diversity function and cluster distribution fall as a power law and, in fact, inversely with i as

$$\langle a_i\rangle=\frac{1}{i} . \tag{2.43}$$

Power-law behavior appears in many areas, some of which are summarized in Mandelbrot [34]. A power-law distribution such as that of Eq. (2.43) has no length scale except for the simple requirement $i<n$. By contrast, a distribution

$$\langle a_i\rangle\sim e^{-i/i_0} \tag{2.44}$$

has a length scale determined by the exponential part and, in particular, the i_0. A way of generating a behavior $1/i$ from e^{-i/i_0} is to have a distribution of length scales present given by some function $f(i_0)$. Then

$$\frac{1}{i}=\int_0^\infty f(i_0)e^{-i/i_0}di_0 , \tag{2.45}$$

where $f(i_0)=1/i_0^2$ for a hyperbolic behavior. All length scales are present since the integral over i_0 is from 0 to ∞.

Power laws in physics signal a critical point [35]. Droplet sizes near a critical point of a liquid-gas phase transition are characterized by a distribution in size which is written as

$$D(k)\sim\frac{1}{k^{\tau}} \tag{2.46}$$

τ is a critical exponent [35], k the size of the drop given by the number of atoms in it, and $D(k)$ the number of drops of size k. Droplet sizes near a critical point are believed to fall with size k with a $\tau\gtrsim 2$ [$\tau=2+(1/\delta)$, where δ is another critical exponent]. For a harmonic series $\tau=1$. Away from the critical point, droplet distributions are postulated to have a behavior given by a scaling relationship [36]

$$D(k)=\frac{1}{k^{\tau}}f(\epsilon^{1/\sigma}k)\ . \tag{2.47}$$

$f(\epsilon^{1/\sigma}k)$ is a scaling function depending on a variable ϵ that is zero at the critical point and is a measure of how far a parameter is away from the critical point. The σ is another critical exponent [36]. The $f(\epsilon^{1/\sigma}k)$ is usually taken to be of the form

$$f(\varepsilon^{1/\sigma}k)=\exp(-\epsilon^{1/\sigma}k)=[z(\varepsilon)]^{k}\ , \tag{2.48}$$

where $z(\epsilon)=e^{-\epsilon^{1/\sigma}}$.

4. *Region $1<\theta<\infty$ and R. A. Fisher's logarithmic series*

The region above $\theta=1$ has a distribution which falls faster than a power. Details can be found in Refs. [2,3]. As an example consider θ an integer, then

$$\langle a_i\rangle=\frac{\theta}{i}\frac{n(n+1)\cdots(n+i-1)}{(\theta+n-1)(\theta+n-2)\cdots(\theta+n-i)}\ . \tag{2.49}$$

For $i\ll n$, $\langle a_i\rangle$ behaves approximately as

$$\langle a_i\rangle\sim\frac{\theta}{i}\left[\frac{n}{n-1+\theta}\right]^{i}\ . \tag{2.50}$$

Defining $y=n/(n-1+\theta)$, the diversity or cluster distribution falls as

$$\langle a_i\rangle=\frac{\langle a_1\rangle}{y}\frac{y^{i}}{i}\ . \tag{2.51}$$

This behavior of $\langle a_i\rangle$ is Fisher's logarithmic series [4]. A discussion of this logarithmic series based on Eq. (2.38) and other situations can also be found in Ref. [8] and references therein. The coefficient $\langle a_1\rangle/y$ in this series is called the coefficient of diversity. In the Fisher analysis [4] of the data of Corbet and Williams, $y=0.997\,428\,1$ and therefore $\langle a_i\rangle$ is nearly hyperbolic and almost without any length scale.

The result of Eq. (2.51) can also be rewritten as

$$\langle a_i\rangle=\frac{1}{\theta^{i-1}}(\langle a_1\rangle)^{i}\frac{1}{i} \tag{2.52}$$

which in cluster descriptions [$\theta=x$, with x given by Eq. (2.10)] has a simple interpretation. The result of Eq. (2.52) is of the form of the law of mass action in chemistry or the Saha equation in astrophysics, as discussed in Refs. [2,3], and is derived under the assumption of chemical equilibrium: $\mu_i=i\mu$.

The above expression can also be thought of as the coalescence [18] of i monomers or singletons into a cluster of size i. The coefficient in this alternate form of Eq. (2.51) is a coalescence probability and this factor appears as $1/\theta^{i-1}$.

F. Moments of cluster distribution and homozygosity

Moments of a distribution of $\langle a_i\rangle$ for $i=1,2,\ldots,n$ are obtained by multiplying each $\langle a_i\rangle$ by $i^{p}, p=1,2,\ldots,$ and taking a sum. The first moment $p=1$ is $\sum_i i\langle a_i\rangle=n$ which follows from the constraint condition. The second moment $p=2$, is [37]

$$\sum_{i=1}^{n} i^{2}\langle a_i\rangle=\frac{n(n+\theta)}{(\theta+1)}\ , \tag{2.53}$$

and the third moment $p=3$ turns out to be

$$\sum_{i=1}^{n} i^{3}\langle a_i\rangle=\frac{n(n+\theta)(2n+\theta)}{(\theta+1)(\theta+2)}\ . \tag{2.54}$$

The result of Eq. (2.53) leads to the simple result

$$\sum_i\frac{i(i-1)}{n(n-1)}\langle a_i\rangle=\frac{1}{\theta+1}\ . \tag{2.55}$$

The expression of Eq. (2.55) has an important interpretation in genetics. Many populations (including humans) are diploid, which means that each individual has two genes at a locus, one from each parent. These two genes will either be of the same allelic type (in which case the individual is said to be a homozygote) or of different allelic types (in which case the individual is a heterozygote). It is well known [27] that in the infinitely many allele model, the mean population homozygosity is $1/\theta+1$. In the context of the samples of n genes, homozygosity can be regarded as the event that two genes taken at random are of the same allelic type. Equation (2.55) shows that the sample probability that this occurs is $1/(\theta+1)$. In other words, the sample statistic $\sum_i i(i-1)a_i/n(n-1)$ is an unbiased estimator of population homozygosity. Clearly, as $\theta\to 0$, the mean homozygosity goes to 1. θ small corresponds to very low mutation rates u, and for small u it becomes increasingly likely that all genes in the sample, and the population, are of the same allelic type. As $\theta\to\infty$, the result of Eq. (2.55) goes to zero and at $\theta=1$, this equation gives $\frac{1}{2}$.

The result of Eqs. (2.53), (2.54), and (2.55) leads to

$$\sum_i\frac{i(i-1)(i-2)}{n(n-1)(n-2)}\langle a_i\rangle=\frac{2}{(\theta+1)(\theta+2)}\ . \tag{2.56}$$

The lhs of Eq. (2.56) represents the probability that three genes in the sample are of the same allelic type. Using diffusion theory results Ewens [17] has shown that the stationary probability that the first i genes are of the same allelic type is given by $(i-1)!\theta/s_i(\theta)\equiv p(i,\theta)$ where $s_i(\theta)=\theta(1+\theta)(2+\theta)\cdots(i-1+\theta)$. For $i=2$, $p(2,\theta)=1/\theta+1$ and for $i=3$, $p(3,\theta)=2/(\theta+1)(\theta+2)$. Similarly $p(4,\theta)=6/(\theta+1)(\theta+2)(\theta+3)$ and

$$\sum_i\frac{i(i-1)(i-2)(i-3)}{n(n-1)(n-2)(n-3)}\langle a_i\rangle=\frac{6}{(\theta+1)(\theta+2)(\theta+3)}\ . \tag{2.57}$$

Sum rules for $\langle a_j a_k \rangle$ are [24]

$$\sum_j \sum_k \frac{kj}{n(n-1)} \langle a_j a_k \rangle = \frac{\theta}{1+\theta} \tag{2.58}$$

and

$$\sum_j \sum_k k^2 j^2 \langle a_j a_k \rangle = \frac{\theta n(n+\theta)[n(n+\theta)-(\theta+1)]}{(\theta+1)(\theta+2)(\theta+3)} . \tag{2.59}$$

G. Maximum entropy methods

The choice of the a_i's which maximize the functional of Eq. (2.15) or the logarithm of the weight function of Eqs. (2.7) or (2.9) subject to the constraint $\sum_i i a_i = n$ can be obtained by the method of Lagrange multipliers. Using Stirling's approximation for the factorials $a_i!$, the resulting a_i's called $\hat{a}_i$'s are

$$\hat{a}_i = \frac{\theta}{i} e^{-\lambda i} . \tag{2.60}$$

The λ is a Lagrange multiplier determined by $\sum_i i\hat{a}_i = n$, which gives

$$n = \frac{\theta z(1-z^n)}{1-z} , \tag{2.61}$$

where $z = e^{-\lambda}$. When $\theta = 1$, the solution of Eq. (2.61) is $z = 1$ and $\hat{a}_i = 1/i$, as before. When $\theta > 1$, $z < 1$ and $z^n \to 0$ for large n. Thus $z = n/n + \theta$ and

$$\hat{a}_i \sim \frac{\theta}{i} \left[\frac{n}{n+\theta} \right]^i \tag{2.62}$$

which again is the same as the Fisher form of Eq. (2.50) for large n. The solution for $\theta < 1$, and therefore $z > 1$, is not as easily obtained. Since $z > 1$, $1 - z^n \sim -z^n$ and z is to be determined by $z^n/(z-1) = n/\theta$. Numerical solutions [37] give results for $\hat{a}_i$ that are reasonably good approximations to the exact solutions for $\langle a_i \rangle$.

When additional constraints are imposed additional Lagrange multipliers have to be included, with one Lagrange multiplier for each constraint. For example, constraining both $n = \sum_i i a_i$ and $m = \sum_i a_i$ leads to a solution of the form

$$\tilde{a}_i = \frac{\theta}{i} e^{-\beta} e^{-\lambda i} = \frac{\theta}{i} e^{-\beta} z^i , \tag{2.63}$$

where $z = e^{-\lambda}$. The two Lagrange multipliers (λ, β) are to be determined by the two constraint equations

$$\frac{n}{\theta} = e^{-\beta} \frac{z(1-z^n)}{1-z} \sim e^{-\beta} \frac{z}{1-z} \tag{2.64}$$

and

$$\frac{m}{\theta} = e^{-\beta} \sum_{i=1}^{n} \frac{z^i}{i} \sim -e^{-\beta} \ln(1-z) . \tag{2.65}$$

The last approximations in these equations are valid when $n \gg 1$, $\theta > 1$, and $z < 1$. Eliminating β gives

$$\frac{n}{m} = \frac{z(1-z^n)}{(1-z)\sum_i z^i/i} = f(z) \sim \frac{-z(1-e^{-(1-z)n})}{(1-z)\ln(1-z)} . \tag{2.66}$$

The z can be determined by plotting $f(z)$ against z, for $z < 1$, and finding that z which gives n/m. The condition $z < 1$ corresponds to $m > \ln n$, since $m \sim \gamma + \ln n$ at $\theta = 1$ and for large n. As an example, for Keith's data given in Ref. [38] the value of $n = 89$ and $m = 15$. The z which gives $n/m = 89/15$ is $z = 0.9458$. For $m \sim \ln n$ and n large the Stirling numbers, which appear in Eq. (2.38), can be approximated by [26] $S_n^m = (n-1)!(\gamma + \ln n)^{m-1}/(m-1)!$. The result of Eq. (2.38) can be written in the following way for $i \ll n$:

$$\langle a_i \rangle_{\pi(n,m)} = \frac{1}{i} \frac{m-1}{\gamma + \ln n} \left[\exp\left[-\frac{m-2}{n(\gamma+\ln n)} + \frac{1}{n} \right] \right]^i . \tag{2.67}$$

For $n = 200$, $m = 10$, the result of the rhs of Eq. (2.66) gives $z = 0.9915$. Equation (2.67) would give a $z = 0.9982$ [from the square brackets in Eq. (2.67)] in good agreement with $z = 0.9915$, considering that a sum was replaced by a logarithm in obtaining $z = 0.9915$. However at large i the small discrepancy in z begins to show in $\langle a_i \rangle$ since z is raised to the ith power in $\langle a_i \rangle$. For $z \approx 1$ $z^i = e^{(1-z)i}$ which is, for $i = 50$, equal to 0.654 at $z = 0.9915$ and 0.914 for $z = 0.9982$. The z^i's are still reasonably close.

The maximum entropy approach is seen to be very useful in easily producing the results of the exact expression. The usefulness of this method in other areas can be found in Ref. [39]. The application of this method to cluster distributions for complex weight factors is discussed in Ref. [22].

H. Cumulative mass distribution, fluctuations, and factorial moments

In Ref. [3], the cumulative mass distribution of cluster sizes and its possible importance was discussed. In the genetic case the quantity $a_j\, j$ is the total number of genes that occur j times. The cumulative distribution

$$M(k) = \sum_{j=1}^{[k]+1} j a_j \tag{2.68}$$

is a staircase function starting at $k = 0$ with $M(1) = a_1$ and ending at $k = n$ with $M(n) = n$. The $[k]$ is the greatest integer in k which is taken to be continuous in $M(k)$. The $M(k)$ for the data of Singh, Lewontin, and Felton [25] is shown in Fig. 3. Every partition has a unique set of steps all ending at $k = n$, $M(n) = n$. There are $p(n)$ possible staircases spanning all different rises and runs, where $p(n)$ is given by Eq. (2.4). Figure 3 shows a long intermission followed by a very large sudden jump.

The ensemble-averaged value of $M(k)$ at $\theta = 1$ is a uniform staircase function since $\langle a_j \rangle = 1/j$. For other θ's, the resulting $M(k)$ can be found. Some examples are in Ref. [3]. The θ hammers the $M(k)$ into various rather smooth homogeneous shapes. When $j a_j$ is characterized by a form $j a_j = \theta e^{-\lambda j}$ and when the sum in Eq. (2.74) is

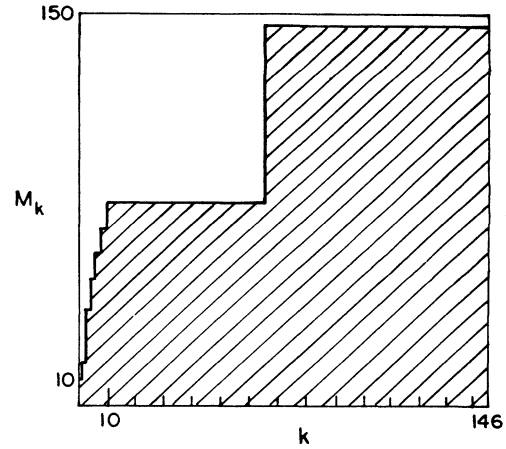

FIG. 3. The cumulative distribution of a partition. The horizontal axis is k and the vertical axis is $M(k)$ defined by Eq. (2.68). The staircase function drawn is for the allelic partition of Fig. 1. The total number of alleles is 146.

changed to an integral, a smoothed cumulative mass distribution called $M_s(k)$ results:

$$M_s(k)=\frac{\theta}{\lambda}(1-e^{-\lambda k}) . \tag{2.69}$$

The difference $M_0(k)-M_s(k)$, where M_0 is the observed $M(k)$, expresses the departure from $M_s(k)$ and $[M_0(k)-M_s(k)]^2$ is its fluctuation.

One method used to study intermittent behavior is based on scaled factorial moments of a distribution [40]. The relevant distribution here is a_j versus j. The j axis $(0,n)$ is divided into bins, not of unit size, but of varying sizes. The largest bin would be the whole interval $(0,n)$. Subsequent division into 2, 3, 4, etc. parts can be made. Let M equal the number of bins giving a bin size $\delta s=n/M$. For example, all alleles in Fig. 1 in a given bin are counted and lumped together to form a quantity $A_J=$ sum of a_i's in bin J. The ith factorial moment is then formed [41] to give

$$\langle F_i\rangle=\frac{\sum_{J=1}^{M}\langle A_J(A_J-1)\cdots(A_J-i+1)\rangle}{\sum_{J=1}^{M}\langle A_J\rangle^i} . \tag{2.70}$$

The $\langle A_J\rangle$ and $\langle F_i\rangle$ are quantities averaged over the possible partitions. A power-law behavior

$$\langle F_i\rangle\sim\left[\frac{n}{\delta s}\right]^{\delta_i} \tag{2.71}$$

may be present in the variation of $\langle F_i\rangle$ with bin size $\delta s=n/M$. The δ_i is called an intermittency exponent and δ_i varies with increasing factorial moment i. The δ_i gives the slope of $\ln\langle F_i\rangle$ versus $\ln\delta s$.

It would be interesting to look for intermittent behavior in allelic distributions. Large fluctuations which seem not to be statistical have been reported in distributions arising from high-energy collisions [41,42]. Turbulent flow patterns may also show intermittency [34].

I. Solutions for more complex weight factors

More complex equilibrium models that go beyond the simple model discussed in Sec. II E can be considered using methods outlined in Sec. II D. The main quantity that has to be determined is $Q_n(\mathbf{x})$ of Eq. (2.23). A solution for the case $\mathbf{x}=(x_1,x_2,\ldots,x_n)=(xy,x,x,\ldots,x)$ or $x_i=xy^{\delta_{i,1}}$ has already been given in Ref. [23]. Here, the procedure used will just be stated. If $x_i=xy^{\delta_{i,k}}$, so that k is given a different weight than the other x_i's which are all equal, then $Q_A(\mathbf{x})$ is to be determined by projection from a generating function:

$$\frac{e^{-x(y-1)u^k/k}}{(1-u)^x}=\sum_n Q_n(\mathbf{x})\frac{u^n}{n!} . \tag{2.72}$$

If two x_i's are different, say, k and j, so that $x_k=xy$ and $x_j=xz$, then the exponential part of (2.72) becomes $-[x(y-1)u^k/k]-x(z-1)u^j/j$. Introducing different x's or θ's would, for example, correspond to having different mutation rates or factors $4N_1u_1,4N_2u_2,\cdots$. Solutions in general are not easily obtained and maximum entropy solutions may again be useful. For example, a simple generalization of Eq. (2.15) to the case of different θ_i's might be

$$\sigma_n(\boldsymbol{\theta})=f(n)\ln\left[\prod_{i=1}^{n}\frac{\theta_i^{a_i}}{a_i!i^{a_i}}\right] . \tag{2.73}$$

The maximum entropy solution is

$$\bar{a}_i=\frac{\theta_i}{i}e^{-\lambda i} \tag{2.74}$$

when the only constraint is $\sum_i ia_i=n$.

J. Equilibrium and nonequilibrium distributions

In several previous sections, mention has been made of chemical potentials and equilibrium constraints. This section is concerned with these quantities. The relations between a_i or n_i and the Lagrange multiplier λ given for example in Eq. (2.60) bear the same formal correspondence between numbers of particles of type k and chemical potentials μ_k that is found in thermodynamics books [28,29,33]:

$$n_k=\frac{V}{v_0}e^{\beta\mu_k} . \tag{2.75}$$

The $\beta=1/k_BT$ and v_0 is given by Eq. (2.11). Spin, mass, binding energy effects, and internal excitation have been neglected to keep the expression simple.

Constraint equations related to particle number conservation in statistical mechanics have a correspondence in thermodynamics with chemical potentials. This relationship is $\lambda_k=\beta\mu_k$ for particles of type k. Chemical equilibrium between various species is expressed by a condition on chemical potentials. Specifically, for a species made of k monomers, the equilibrium condition is $\mu_k=k\mu_1$, where μ_1 is the chemical potential of the monomer. The

exponential part of Eq. (2.75) is then $e^{\beta\mu_k}=e^{\beta\mu_1 k}=e^{-\lambda k}$, with $\lambda=-\beta\mu_1$. The factor $e^{-\lambda k}$ appears in Eq. (2.60) and was obtained by maximizing the entropy with respect to variations of the n_k's (or a_k's). Thus a generalization of Eq. (2.74) is to allow $a_i=(\theta_i/i)e^{-\lambda_i}$ where $\lambda_i\neq i\lambda$ in general. If equilibrium is present, then $\lambda_i=i\lambda$.

The entropy S, Gibbs potential G, internal energy U, and Helmholtz free energy F each contain parts [28] that involve μ_k and n_k as $\sum_k\mu_k n_k$ and variations that contain $\sum_k\mu_k\delta n_k$, so that $\delta S\sim\sum_k-\beta\mu_k\delta n_k$ as an example. In fact G is exactly $\sum_k\mu_k n_k$. Since $n=\sum_k kn_k$, variations in the n_k's are not independent, but must satisfy the constraint $0=\sum_k k\delta n_k$. Again $P\equiv S$, U, F, or G is stationary with respect to changes in the n_k's when the μ_k's satisfy $\mu_k=k\mu_1$. The constraint can be incorporated into P by forming a function h: $h=P-\lambda'(1-\sum_k kn_k)$. This h is stationary with regard to variations of the δn_k's when $\delta h/\delta n_k=0$ for each k, which gives $\mu_k=\lambda' k$; since $\mu_1=\lambda'$, $\mu_k=k\mu$ as before.

Away from an equilibrium point, S is not a maximum; also G, U, and F are not at their minima. The n_k's will change in such a way as to increase S or decrease G, U, and F. As a very simple example, consider a two-component system with $N=n_a+n_b$. $G=\mu_a n_a+\mu_b n_b$. Since $\delta n_a=-\delta n_b$, δG is given by $\delta G=(\mu_a-\mu_b)\delta n_a$. If $\mu_a>\mu_b$, then n_a's must decrease or $\delta n_a<0$ to lower G. The a's are changed into b's. If $\mu_a<\mu_b$, the same type of argument shows b's$\rightarrow a$'s to lower G. Equilibrium is reached when $\mu_a=\mu_b$. The μ's act as generalized forces or pressures which lead to transformations between species.

As another example, consider the simple reaction which transforms k monomers into one cluster of size k. The $k\,\delta n_1=-1\delta n_k$, where k and 1 are the coefficients of the reaction called stochiometric coefficients in chemistry and are given the symbol ν. In general $\delta n_i=\nu_i\delta x$ with δx an arbitrary change. The ν_i's can be positive, for reactants, or negative, for products. The variation for G for the simple case $k\delta n_1=-\delta n_k$ is

$$\delta G=(\mu_k-k\mu_1)\delta n_k\ . \tag{2.76}$$

An equilibrium equality between chemical potentials also appears in phase transitions such as from a liquid to a gas. At equilibrium $\mu_L=\mu_G$ where μ_L, μ_G are the liquid and gas chemical potentials, respectively. The Gibbs criteria for phase equilibria are equality of the temperatures, $T_L=T_G$, equality of the pressures, $P_L=P_G$, and equality of the chemical potentials, $\mu_L=\mu_G$.

The role of the $\beta\mu$'s as "generalized forces or pressures" will now be considered. A process involving two clusters A,B which interact to make two other clusters C,D is

$$A+B\rightleftarrows C+D\ . \tag{2.77}$$

In the above reaction, the stoichiometric coefficients are taken as unity for simplicity. Simple kinetic theory arguments [33] will be used to investigate the transformations of $A+B$ into $C+D$ and vice versa. The formation of C, for example, is proportional to the number of A's and B's. From left to right in the above equation, the increase in C per unit time is $\dot{n}_C^{(+)}=K_+ n_A n_B$. K_+ is the proportionality constant. However, the back process from right to left decreases C and this decrease is proportional to the number of C's and D's. Specifically, $\dot{n}_C^{(-)}=K_- n_C n_D$ where K_- is the proportionality constant for the back process. The total change in C is then $\dot{n}_C=K_+ n_A n_B-K_- n_C n_D$. At equilibrium $\dot{n}_C=0$ and $K^+/K^-=\tilde{n}_C\tilde{n}_D/\tilde{n}_A\tilde{n}_B$ where the $\tilde{n}$'s are the equilibrium numbers of A, B, C and D. Using the results of Eq. (2.75) the following equation is obtained:

$$\dot{n}_C=K_+ n_A n_B(1-e^{-\beta A_c})\ , \tag{2.78}$$

where the chemical activity A_c is

$$A_c=(\mu_A+\mu_B)-(\mu_C+\mu_D)\ . \tag{2.79}$$

When $\mu_A+\mu_B=\mu_C+\mu_D$, $\dot{n}_C=0$ and n_C is in equilibrium. The constraint $\mu_A+\mu_B=\mu_C+\mu_D$ is just the statement of chemical equilibrium for Eq. (2.77).

The relationship in Eq. (2.78) between $\dot{n}_C$ and A_c is nonlinear. When A_c is small $1-e^{-\beta A_c}=1-(1-\beta A_c)=\beta A_c$ and

$$\dot{n}_C=K_+ n_A n_B\beta A_c\ . \tag{2.80}$$

Then a linear relationship results in this approximation. Thus, in both cases changes in n_C result from a driving term or generalized force term involving the activity A_c. If $\mu_C+\mu_D<\mu_A+\mu_B$ then A_c is positive, $\dot{n}_C$ is positive, and C increases since $C+D$ has the lower chemical potential. If $\mu_C+\mu_D>\mu_A+\mu_B$, then A_c is negative, $\dot{n}_C$ is negative, and C decreases since $A+B$ has the lower chemical potential.

The simple linear relationship of Eq. (2.80) is of the form of Ohm's law $J=\sigma E$, where J is the current density, σ the conductivity, and E the applied electric force field.

Another classic situation is a droplet in a vapor. A nonequilibrium situation arises when a vapor is supersaturated. Classical nucleation theory is based on the following picture [43,21]. A drop is treated as a liquid with chemical potential μ_L while a vapor has a chemical potential μ_V. Let N_L be the number of atoms in a drop and N_V be the number of atoms in a vapor. Then, to lowest order, the difference in Gibbs's potentials with and without the drop is

$$\Delta G=\mu_L N_L+\mu_V N_V-\mu_V(N_L+N_V)=(\mu_L-\mu_V)N_L\ . \tag{2.81}$$

To this result a term is added which arises from the creation of an interfacial surface between the drop and the vapor:

$$\Delta G=(\mu_L-\mu_V)N_L+4\pi R_L^2\sigma\ , \tag{2.82}$$

where σ is a surface tension and R_L the radius of the liquid drop. The radius R_L and N_L are related by the number density $\rho_L=N_L/(4\pi R_L^3/3)$ in the liquid drop. The evaporation, metastability, and growth of a drop are governed by Eq. (2.82). A drop is metastable when $d\Delta G/dR_L=0$. The R_L which satisfies this condition is

$$R_L^*=\frac{2\sigma}{(\mu_V-\mu_L)\rho_L} . \tag{2.83}$$

Smaller drops $R_L<R_L^*$ evaporate atoms while larger drops $R_L>R_L^*$, grow by accumulation of vapor atoms to form still larger drops. In both cases, growth and evaporation, ΔG is lowered by the processes. Without the extra surface term in Eq. (2.82), $\Delta G=(\mu_L-\mu_V)N_L$ and a drop grows for any R_L if $\mu_L-\mu_V<0$ or evaporates for any R_L if $\mu_L>\mu_V$. The runaway over critical drop threatens the existence of the vapor phase. The end product of a supersaturated system with an overcritical drop is a liquid-vapor phase separation. Below a temperature T_c, called a critical temperature, a liquid and vapor can coexist.

The evolution of a charged drop in a supersaturated system appears in the theory of cloud chambers. In the case of charged drops, even small drops can grow when $\Delta_\mu=\mu_V-\mu_L$ exceeds a certain threshold. For uncharged drops the change in ΔG with N_L is given by

$$\frac{d\Delta G}{dN_L}=\mu_L-\mu_V+\frac{2\sigma}{\rho_L R_L} . \tag{2.84}$$

The metastable line is determined by $d\Delta G/dN_L=0$, which gives Eq. (2.83). Regions of droplet growth correspond to $d\Delta G/dN_L<0$ and droplet evaporation corresponds to $d\Delta G/dN_L>0$. For a charge drop [44], Eq. (2.84) acquires an additional term:

$$\frac{d\Delta G}{dN_L}=\mu_L-\mu_V+\frac{2\sigma}{\rho_L R_L}-\frac{\kappa-1}{\kappa}\frac{1}{8\pi}\frac{q_e^2}{\rho_L R_L^4} , \tag{2.85}$$

where κ is the dielectric constant of the liquid drop and q_e is the charge on the drop. The metastability line is determined by $d\Delta G/dN_L=0$. Figure 4 illustrates the charged and uncharged cases.

Allelic distributions are stationary (Hardy-Weinberg equilibrium) in the absence of mutation, random drift, migration, and natural selection [27]. When these forces are absent, random mating leaves the frequency of occurrence of different alleles unchanged, a result which is called the Hardy-Weinberg law. However, if one allele has, for example, a selective advantage over others, then random mating produces allelic frequency changes. Such differential fitnesses between alleles results in nonequilibrium situations which break this Hardy-Weinberg equilibrium. These differential fitnesses act as pressures or generalized forces acting in a manner similar to chemical potentials in cluster changes. Differential fitnesses lead to the growth in frequency of occurrence of some alleles and the disappearance of others. As an illustration consider a case of two alleles A and a with diploids AA, Aa, and aa having fitnesses $(1+s)^2$, $1+s$, and 1, and probabilities of occurrence p^2, $2pq$, and q^2 respectively. After one random mating the initial frequency $f_A(0)$ of A changes to $f_A(1)$ given by [27]

$$f_A(1)=\frac{f_A(0)(1+s)}{f_A(0)(1+s)+1-f_A(0)} .$$

After t generations, $f_A(t)$ is

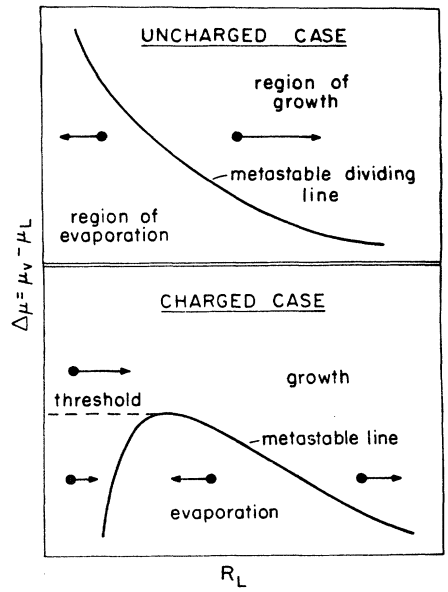

FIG. 4. Regions of growth and evaporation for uncharged and charged drops. The vertical axis is $\Delta\mu=\mu_V-\mu_L$ and the horizontal axis is the radius of the drop R_L. The metastable line is determined by Eq. (2.83) for the uncharged case and by Eq. (2.84) for the charged case. For a given $\Delta\mu$, growing drops are characterized by a line pointing to the right while evaporating drops are characterized by a line pointing to the left.

$$f_A(t)=\frac{f_A(0)(1+s)^t}{f_A(0)(1+s)^t+1-f_A(0)} .$$

For $s>1$, $f_A(t)\to 1$ and $f_a(t)=1-f_A(t)\to 0$ as $t\to\infty$. For $s<1$, $f_A(t)\to 0$ and $f_a(t)=1-f_A(t)\to 1$ as $t\to\infty$.

When fitnesses are $(1-s_A)$ for AA, 1 for Aa, and $(1-s_a)$ for aa, equilibrium points exist at $f_A=0,1$ and $s_a/(s_A+s_a)$. When s_a and s_A are negative, the heterozygote Aa has the lowest fitnesses and is underdominant. The value $s_a/(s_A+s_a)$ corresponds to an unstable equilibrium point. For $f_A(0)>s_a/(s_A+s_a)$, $f_A(t)\to 1$ as $t\to\infty$ and for $f_A(0)<s_a/(s_A+s_a)$, $f_A(t)\to 0$ and $t\to 0$. This behavior in f_A is similar to a drop in a supersaturated vapor. By contrast, the overdominant case has s_a, s_A positive. The $s_a/(s_A+s_a)$ is a stable equilibrium point and this situation has no correspondence with a drop in a supersaturated vapor.

III. PROBABILITY CONCEPTS IN CLUSTER AND GENETIC DIVERSITY DISTRIBUTIONS

This section is concerned with probability concepts associated with the distribution $\langle a_k\rangle$ given by Eq. (2.35). Since $\sum_k k\langle a_k\rangle=n$, a quantity

$$p_n(k,\theta)=\frac{k\langle a_k\rangle}{n} \tag{3.1}$$

defines the fraction of the mass in clusters of size k or a fraction of alleles occurring k times. The $p_n(k,\theta)$ satisfy $\sum_k p_n(k,\theta)=1$, $1\geq p_n(k,\theta)\geq 0$, and therefore $p_n(k,\theta)$ can

be considered a probability function. Using Eq. (2.35)

$$p_n(k,\theta)=\begin{bmatrix}n-1\\k-1\end{bmatrix}\theta B(\theta+n-k,k)\ , \tag{3.2}$$

where the $B(\theta+n-k,k)$ is a beta function: $B(w,z)=\Gamma(w)\Gamma(z)/\Gamma(w+z)$ where $\Gamma(w)$ is a gamma function.

A. Allelic or cluster distribution $k\langle a_k\rangle/n$ and Bernoulli trials; R. A. Fisher's law

The $p_n(k,\theta)$ of Eq. (3.2) can be rewritten as

$$p_n(k,\theta)=\int_0^1\begin{bmatrix}n-1\\k-1\end{bmatrix}p^{k-1}(1-p)^{n-k}u(\theta,p)dp \tag{3.3}$$

using an integral representation of the beta function. $u(x,p)$ is

$$u(\theta,p)=\theta(1-p)^{\theta-1}\ . \tag{3.4}$$

The part $\binom{n-1}{k-1}p^{k-1}(1-p)^{n-k}$ corresponds to Bernoulli trials with probability p of heads. The shift by one arises because success varies as $0,1,2,\ldots$ while $k=1,2,3,\ldots$ The $\int_0^1 u(\theta,p)dp=1$ and $u(\theta,p)$ is a probability density function. When $\theta=0$, $u(\theta,p)=\delta(1-p)$ where $\delta(1-p)$ is a Dirac delta function at $p=1$. Then $p_n(k,\theta=0)=1$ for $k=n$, otherwise 0. When $\theta=1$, $u(\theta,p)=1$ and $u(\theta,p)$ is a uniform distribution giving $p_n(k,\theta=1)=1/n$ for $k=1$ to n. When $\theta\to\infty$, $u(\theta,p)=\delta(p)$ and is a delta function at $p=0$ or zero probability for heads. Consequently $p_n(k,\theta=\infty)=1$ for $k=1$, otherwise zero. The uniform case also follows from the Bayes argument as given in Feller [45].

The $p_n(k,\theta)$ is a randomized Bernoulli distribution obtained from a mixed population of coins with a distribution of p's given by the $u(\theta,p)$ of Eq. (3.4). The resulting $\langle a_k\rangle=np_n(k,\theta)/k$ is Eq. (2.35). Section II E contains a discussion of the properties of $\langle a_k\rangle$ and, in particular, Sec. II E 4 gives a logarithmic series form for $\langle a_k\rangle$.

R. A. Fisher developed the logarithmic series for species diversity by a different procedure, namely, a Poisson distribution

$$p(n)=e^{-m}\frac{m^n}{n!} \tag{3.5}$$

was randomized with a Eulerian or χ^2 density distribution:

$$u(m,p,k)dm=\frac{1}{(k-1)!}p^{-k}m^{k-1}e^{-m/p}dm\ . \tag{3.6}$$

Here m is the expected number of species and $p(n)$ the probability of observing n as determined by a Poisson distribution. The value of m is then treated as a variable with a density $u(m,k,p)$ where k and p are parameters. The p is proportional to the size of the sample and k measures the variations in m and gives a mean value for $m=pk$. The resulting integral is

$$p(n;p,k)=\frac{(k+n-1)!}{(k-1)!n!}\frac{p^n}{(1+p)^{k+n}} \tag{3.7}$$

and is related to a negative binomial. A logarithmic series follows upon further approximations (neglecting zeros and for small k parameter).

It should be noted that the weight function of Eqs. (2.7) or (2.9) can be recast into a form that looks like multiple Poisson distributions. Writing $y=e^{-\beta}=n/(n-1+\theta)$ and using the approximate form of Eq. (2.50) for $\langle a_i\rangle$,

$$p(\mathbf{a},\theta)\sim\left[\Pi_i\frac{\langle a_i\rangle^{a_i}}{a_i!}e^{-\langle a_i\rangle}\right]\frac{e^{\beta n}e^{\langle m\rangle}n!}{L(\theta)}\ . \tag{3.8}$$

The $\langle m\rangle=\langle a_1+a_2+\cdots+a_n\rangle$ is the mean number of alleles or mean multiplicity of clusters. The $\langle m\rangle$ is

$$\langle m\rangle=\theta\left[\frac{1}{\theta}+\frac{1}{\theta+1}+\cdots+\frac{1}{\theta+n-1}\right] \tag{3.9}$$

as was found by Ewens in the allele case [1] and the same result was found in the cluster case [3].

B. Allelic or cluster distribution $k\langle a_k\rangle/n$ as the Pólya-Eggenberger distribution

The $p_n(k,\theta)$ can be obtained by replacement sampling from an urn. Consider an urn with γ red balls and ω white balls. With each drawing of a ball, s balls, of the same color drawn, plus the original ball, are returned. The probability of m successful red drawings in n' trials is the Pólya-Eggenberger distribution [45,46]:

$$P(m,\alpha,\beta)=\begin{bmatrix}n'\\m\end{bmatrix}\frac{B(\alpha+m,\beta+n'-m)}{B(\alpha,\beta)}\ , \tag{3.10}$$

where $\alpha=r/s$, $\beta=w/s$, and the B's are beta functions already defined. The result of Eq. (3.2) is obtained when $n'=n-1$, $m=k-1$, $\alpha=1$, and $\beta=\theta$. When $\alpha=1$, $r=s$, and $\theta=w/r$. At $\theta=1$, $w=r=s$, and $i\langle a_i\rangle=1$, $p_n(i,\theta=1)=1/n$. A different urn was considered by Hoppe [15] and Donnelly [16] for the Ewens sampling theory.

In the Pólya-Eggenberger distribution the occurrence of something (drawing red) increases the chance of its occurrence on the next trial. For example, if the first m balls drawn are red the chance of drawing a red ball on the next trial is

$$\frac{m}{m+\theta} \tag{3.11}$$

when $r=s$ or $\alpha=1$ and $\theta=w/r$. After m drawings of red, the urn contains $r+(m-1)r=mr$ red balls and w white balls and the above result easily follows. The probability of drawing a white ball is then

$$\frac{\theta}{m+\theta}\ . \tag{3.12}$$

It should be noted that the probability of a given sequence, say, *rrrrww*, is the same as any other rearrangement of the r's and w's as can easily be verified. The binomial factor of Eq. (3.10) just counts all the possible arrangements of mr's and $(n'-m)w$'s.

The maximum entropy solution of Eq. (2.60) turns out to be Jayne's dice model [47]. A die is rolled and the only information known after a large number of rolls is the mean number of dots. Given the value for the mean, how

should the probability for each side $p(k)$ be determined? Jayne's model maximizes $-\sum_k p(k)\ln p(k)$ subject to the constraint imposed by a specification of the mean number of dots.

IV. SUMMARY

In this paper a parallel is drawn between theories of cluster distributions in physics and issues related to genetic diversity in biology. Specifically, a recent simple model for cluster distributions is shown to have a formal mathematical structure very similar to the Ewens sampling theory of genetic diversity. In fact, a simple transcription of terms connects these two areas. This transcription related the number of clusters n_k of size k to the number of different alleles a_k each appearing k times. The mutation rate in genetics has its analog in Richardson's thermionic emission rate in this correspondence. Properties of the distribution of cluster sizes are related to various quantities which appear in genetic diversity. For example, the second moment of the cluster distribution is related to a quantity called the homozygosity in genetics. The homozygosity is a measure of genetic similarity in an allelic distribution.

Because of this possible connection, methods from statistical mechanics and kinetic theory which have been used in theories of cluster distributions may also be carried over into a discussion of genetic diversity. Maximum "entropy" methods are shown to give very good approximate solutions for exactly soluble models for both cluster and allelic distributions. Moreover, such methods may then be useful in discussing more complex models which are not exactly soluble. A larger class of approximate solutions can then be obtained from these more complex models.

A solution to a simple model discussed in this paper is also shown to contain R. A. Fisher's logarithmic distribution as an approximation. This distribution expresses the biological diversity of different species as a logarithmic power series. Cluster distribution based on chemical equilibrium laws are also discussed and nonequilibrium features are mentioned in an attempt to go beyond equilibrium distributions. Examples of some nonequilibrium situations are given. In the cluster case, these examples are drops in a supersaturated vapor and reactions driven by chemical potential differences. A supersaturated vapor also has a chemical potential difference between a vapor and a drop. These chemical potential differences act as generalized forces or pressures changing the cluster distributions. An analogous situation for allelic distribution arises when one type of allele has a selective advantage over other alleles. Then this allele can grow in frequency of occurrence in random matings based on Mendelian laws.

Finally, it should be emphasized that the formal correspondence between cluster and allelic distribution does not imply some similarity in the basic underlying processes. The dynamical processes which lead to the biological diversity and to the fragmentation distributions are not simply related to one another. However, drawing this parallel may lead to new insights into each field derived from the other field. Moreover, the recognition that similar mathematical descriptions can be employed in fragmentation physics and genetics does create the possibility that methods developed in one area may also be used with some advantage in the other area.

ACKNOWLEDGMENTS

The author would like to thank Joe Felsenstein for introducing him to topics in population genetics, and Warren Ewens for very helpful discussions. This work was supported in part by a grant from the National Science Foundation, Grant No. NSF-89-03457, and in part by the U.S. Department of Energy.

APPENDIX

A heuristic argument for why Eq. (2.9) contains k^{n_k} and not $(k!)^{n_k}$ is as follows. The internal partition function of a cluster of size k is proportional to $V_k^k/k!$ with $k!$ arising because particles in V_k are identical. The V_k is the volume of the cluster which is taken to be proportional to the number of particles contained in the volume: $V_k \sim v_0 k$. The resulting k^k from V_k^k removes the factorial dependence in favor of a simple power k^α, where α is a number of the order of 1. When n_k clusters of size k are present, this result is raised to the n_kth power. A more general weight would involve $\pi_k(k^\alpha)^{n_k} n_k!$ in the denominator of Eq. (2.9) and $L(x)$ would be different. The case $\alpha=1$ is easily solved. The factor $\pi_k(k!)^{n_k} n_k!$ would arise in situations in which V_k is independent of k.

[1] W. J. Ewens, Theor. Popul. Biol. **3**, 87 (1972).
[2] A. Z. Mekjian, Phys. Rev. Lett. **64**, 2125 (1990).
[3] A. Z. Mekjian, Phys. Rev. C **41**, 2103 (1990).
[4] R. A. Fisher, A. S. Corbet, and C. B. Williams, J. Animal Ecol. **12**, 42 (1943).
[5] A. Sommerfeld, *Thermodynamics and Statistical Mechanics* (Academic, New York, 1956).
[6] S. Karlin and J. L. McGregor, Theor. Popul. Biol. **3**, 113 (1972).
[7] G. A. Watterson, Adv. Appl. Probab. **6**, 463 (1974).
[8] D. A. Watterson, Theor. Popul. Biol. **6**, 217 (1974).
[9] A. C. Trajstman, Adv. Appl. Probab. **6**, 489 (1974).
[10] W. J. Ewens, and J. H. Gillespie, Theor. Popul. Biol. **6**, 35 (1974).
[11] K. Kirby, Theor. Popul. Biol. **7**, 277 (1975).
[12] J. F. C. Kingman, Theor. Popul. Biol. **11**, 274 (1977).
[13] R. C. Griffiths, Adv. Appl. Probab. **11**, 326 (1979).
[14] W. J. Ewens, *Mathematical Population Genetics* (Springer-Verlag, Berlin, 1979).
[15] F. M. Hoppe, J. Math. Biol. **20**, 91 (1984).
[16] P. Donnelly, Theor. Popul. Biol. **30**, 271 (1986).
[17] W. J. Ewens, *Population Genetics Theory—the Past and the Future in: Mathematical and Statistical Developments in Evolutionary Theory,* edited by S. Lessard (Kluwer

Academic, Dordrecht, 1990), pp. 177–227.
[18] A. Z. Mekjian, Phys. Rev. C **17**, 1051 (1978).
[19] A. Z. Mekjian, Nucl. Phys. A **384**, 492 (1982).
[20] H. Jaqaman, A. Z. Mekjian, and L. Zamick, Phys. Rev. C **27**, 2782 (1983).
[21] A. Goodman, J. Kapusta, and A. Z. Kekjian, Phys. Rev. C **30**, 851 (1984).
[22] A. R. De Angelis and A. Z. Mekjian, Phys. Rev. C **40**, 105 (1989).
[23] S. J. Lee, and A. Z. Mekjian, Phys. Lett. **149** 7 (1990).
[24] S. J. Lee and A. Z. Mekjian, Phys. Rev. A **44**, 6294 (1991).
[25] R. S. Singh, R. C. Lewontin, and A. A. Felton, Genetics **84**, 609 (1976).
[26] *Handbook of Mathematical Functions,* edited by M. Abramowitz and I. Stegun, Natl. Bur. Stand. Appl. Math. Ser. No. 55 (U. S. GPO, Washington, DC, 1965).
[27] D. L. Hartl and A. G. Clark, *Principles of Population Genetics* (Sinauer, Sunderland, MA, 1989).
[28] P. Morse, *Thermal Physics* (Benjamin/Cummings, Reading, MA, 1969).
[29] A. Sommerfeld, *Thermodynamics and Statistical Mechanics* (Academic, New York, 1956).
[30] J. Riordan, *An Introduction to Combinatorial Analysis* (Wiley, New York, 1958).
[31] M. Hammermesh, *Group Theory and its Applications to Physical Problems* (Addison-Wesley, Reading, MA, 1962).
[32] J. F. C. Kingman, *Mathematics of Genetic Diversity* (SIAM, Philadelphia, 1980).
[33] E. Fermi, *Thermodynamics* (Dover, New York, 1956).
[34] B. B. Mandelbrot *The Fractal Geometry of Nature* (Freeman, San Francisco, 1982).
[35] M. E. Fisher, Physics **3**, 255 (1967).
[36] D. Stauffer, *Introduction to Percolation Theory* (Taylor and Francis, Philadelphia, 1985).
[37] A. Mekjian, Rutgers University Report No. RU90-07 (unpublished).
[38] T. P. Keith, L. D. Brooks, R. C. Lewontin, J. C. Martinez-Cruzado, and D. L. Rigby, Mol. Biol. Evol. **2**, 206 (1985).
[39] *Maximum Entropy and Bayesian Methods in Applied Statistics,* edited by J. H. Justice (Cambridge University Press, Cambridge, 1984).
[40] A. Bialas and R. Peschanski, Nucl. Phys. B **273**, 703 (1986).
[41] M. Ploszajczak, and A. Tucholski, Phys. Rev. Lett. **65**, 1539 (1990).
[42] R. Holynski, *et al.* Phys. Rev. Lett. **62**, 733 (1989).
[43] *Nucleation,* edited by A. C. Zettlemoyer (Dekker, New York, 1969).
[44] J. D. Wilson, *The Principles of Cloud Chamber Techniques* (Cambridge University Press, Cambridge, 1951).
[45] W. Feller, *An Introduction to Probability Theory and its Applications,* VI,2 (Wiley, New York, 1957).
[46] N. Johnson, and S. Koltz, *Discrete Distributions* (Houghton Mifflin, Boston, 1969).
[47] E. T. Jayne, *Where Do We Stand on Maximum Entropy in the Maximum Entropy Formalism,* edited by D. Levine and M. Tribus (MIT Press, Cambridge, MA, 1979).