A GENEALOGICAL DESCRIPTION OF THE INFINITELY-MANY NEUTRAL ALLELES MODEL

P. J. Donnelly
Department of Statistical Science
University College London
London WC1E 6BT
ENGLAND

S. Tavaré
Department of Mathematics
University of Utah
Salt Lake City, UT 84112
U.S.A.

1. INTRODUCTION

Kingman [8], [9] introduced the coalescent as a means of describing the genealogy of samples taken from a large evolving haploid population. The coalescent partitions a sample of genes into equivalence classes with respect to an ancestral population some time t into the past; genes in the same equivalence class in the sample have the same ancestors. As t increases the equivalence classes coalesce until, sufficiently far in the past, all individuals in the sample are equivalent, being descended from a common ancestor.

Watterson [11] showed that the effects of mutation in the genealogy could be allowed for explicity by constructing a process with two different types of equivalence classes. Specifically, consider a sample of n genes chosen at reference time 0. Genes i and j are in the same "old" equivalence class at time $-t$ if i and j share a common ancestor at time $-t$, and no mutation has occurred in the line of descent from that ancestor to i and j between time 0 and time $-t$. On the other hand, i and j might be descended from a common ancestor at time $-s$ $(> -t)$, that ancestor being a new mutant. If no further mutation occurs in the line of descent to i and j, then we say that i and j are in the same "new" equivalence class. Each new equivalence class contains genes of identical type, and with the infinitely-many alleles assumption that each mutation leads to a novel allelic type, distinct new equivalence classes have distinct types.

Donnelly and Tavare [3] observed that keeping track of the new equivalence classes in order of their appearance leads to a direct way of studying questions involving the ages of alleles and the age ordering in samples.

In this paper, we will focus on an age-ordered genealogical description of the infinitely-many neutral alleles diffusion model from which the samples were taken. After a brief review of the stochastic structure of the age-ordered sample coalescent, we construct a Markov process which keeps track of the allele frequencies in the (infinite) population and the order of their occurrence by mutation. In the final section, we use this population frequency process to give a unified treatment of several aspects of neutral mutation theory.

2. A COALESCENT WITH AGES

We will need later the basic properties of the sample coalescent with ages. The necessary results are taken, with minor changes of notation, from [3]. This process is a continuous-time Markov chain $\{A_t, t \geq 0\}$ in which a typical state can be represented as a collection $(\xi_1, \ldots, \xi_k; \eta_1, \ldots, \eta_\ell)$ of equivalence classes in which $\xi_1, \ldots, \xi_k$ denote old classes, and $\eta_1, \ldots, \eta_\ell$ represent new classes, those genes in class η_r having arisen by mutation after (that is, further into the past!) those in class η_s if $r < s$. Let $\mathcal{E}_n^*$ be the collection of such ordered equivalence relations. Initially, each individual is in his own old equivalence class, and there are no new classes. For sufficiently large t, all individuals in the sample are in new classes, since the sample must comprise only mutants with respect to sufficiently remote generations.

One of the attractive features of these coalescents is that the process changes state through one of only two possible mechanisms: either two old equivalence classes coalesce to form a single (old) class, or an old class becomes the oldest of the new classes. In each case, the number of old classes decreases by exactly one, and the times between such changes depend only on the number of old classes then present, and not on the detailed structure of the current configuration.

Let $\{D_t, t \geq 0; D_0 = n\}$ be a pure death process on $\{n, n-1, \ldots, 0\}$ with death rate $k(k+\theta-1)/2$ from state k, where $\theta > 0$ is the scaled mutation parameter. Let $\{\mathcal{Q}_k, k = n, \ldots, 0\}$ be a discrete-time Markov chain on $\mathcal{E}_n^*$, independent of the death process. $\mathcal{Q}_k$ is an element of $\mathcal{E}_n^*$ having k old equivalence classes, and the transition probabilities are given by

$$P(\mathcal{A}_{k-1}=(\xi_1,\ldots,\xi_{i-1},\xi_i,\ldots,\xi_k;\xi_i,\eta_1,\ldots,\eta_\ell)\,|\,\mathcal{A}_k=(\xi_1,\ldots,\xi_k;\eta_1,\ldots,\eta_\ell)) = \theta/k(k+\theta-1). \tag{2.1a}$$

$$P(\mathcal{A}_{k-1}=(\xi_1,\ldots,\xi_i\cup\xi_j,\ldots,\xi_k;\eta_1,\ldots,\eta_\ell)\,|\,\mathcal{A}_k=(\xi_1,\ldots,\xi_k;\eta_1,\ldots,\eta_\ell)) = 2/k(k+\theta-1), \tag{2.1b}$$

for $k = 1,\ldots,n$; $1 \leq i < j \leq k$. Here $\xi_i \cup \xi_j$ denotes the union of the two old equivalence classes ξ_i and ξ_j. The marginal distribution of $_k$ is obtained from [3], equation (3.2):

$$P(\mathcal{A}_k=(\xi_1,\ldots,\xi_k;\eta_1,\ldots,\eta_\ell)) = \frac{\Gamma(k+\theta)(n-k)!k!\theta^\ell\lambda_1!\lambda_2!\ldots\lambda_k!\mu_1!\ldots\mu_\ell!}{n!\Gamma(n+\theta)(\mu_1+\ldots+\mu_\ell)(\mu_2+\ldots+\mu_\ell)\ldots(\mu_{\ell-1}+\mu_\ell)\mu_\ell}, \tag{2.2}$$

where λ_i is the number of genes in ξ_i and μ_i is the number of genes in η_i. Finally, we have

$$A_t = \mathcal{A}_{D_t}, \quad t > 0. \tag{2.3}$$

The results (2.2) and (2.3) provide a detailed description of age-ordered samples from the infinitely-many neutral alleles model. A number of further applications are given in [3]. We turn now to the corresponding population structure.

3. THE POPULATION FREQUENCY PROCESS

If we are not particularly interested in which genes in our sample belong to which equivalence classes, but rather in the numbers of genes in those classes, we obtain another Markov process $\{M_t, t>0\}$ whose structure follows immediately from (2.1)-(2.3). Its jump chain, $\mathcal{M}_k$, consists of collections of integers of the form $(\lambda_1,\ldots,\lambda_k;\mu_1,\ldots,\mu_\ell)$ giving the numbers of individuals in each of the equivalence classes of $\mathcal{A}_k$, and

$$P(\mathcal{M}_k=(\lambda_1,\ldots,\lambda_k;\mu_1,\ldots,\mu_\ell)) = \frac{\Gamma(k+\theta)(n-k)!k!\theta^\ell}{\Gamma(n+\theta)\alpha_1!\alpha_2!\ldots\alpha_n!\mu_\ell(\mu_\ell+\mu_{\ell-1})\ldots(\mu_\ell+\ldots+\mu_1)} \tag{3.1}$$

where $\alpha_i = \#\{j:\lambda_j=i\}$. See [11] equation (3.3.1) for the corresponding result where age ordering is ignored.

In order to describe the population structure that corresponds, as it were, to samples of size $n = \infty$, we will need the following notation. Let $\mathbb{N}_0 = \{0,1,\ldots,\infty\}$, let Δ be the collection of sequences $\{x_1, x_2, \ldots\}$ satisfying

$$x_i \geq 0, \ \sum_{i=1}^{\infty} x_i \leq 1,$$

and Δ^1 the subset of Δ comprising those sequences with sum 1. Let $S = \mathbb{N}_0 \times \Delta$. Fix $k \geq 1$, and let V_k be a random variable having the beta density f_k given by

$$f_k(x) = \frac{\Gamma(k+\theta) x^{k-1} (1-x)^{\theta-1}}{\Gamma(\theta)(k-1)!}, \ x \in (0,1).$$

Let $(U_1,\ldots,U_k)$ be a random k-vector having a uniform distribution on the simplex $\{(u_1,\ldots,u_k); u_i \geq 0, u_1+u_2+\ldots+u_k = 1\}$ (i.e. density proportional to Lebesgue measure on the simplex), and let Z_1, $Z_2, \ldots$ be independent identically distributed random variables with density f given by

$$f(x) = \theta(1-x)^{\theta-1}, \ x \in (0,1), \tag{3.2}$$

and take $\{U_i\}$, V_k and $\{Z_j\}$ mutually independent. Finally, define the random vector $\underset{\sim}{F}^{(k)}$ by

$$\begin{aligned}\underset{\sim}{F}^{(k)} = (&V_k U_1, V_k U_2, \ldots, V_k U_k, (1-V_k) Z_1, \\ &(1-V_k)(1-Z_1) Z_2, (1-V_k)(1-Z_1)(1-Z_2) Z_3, \ldots).\end{aligned} \tag{3.3}$$

When $k = 0$, put

$$\underset{\sim}{F}^{(0)} = (Z_1, (1-Z_1)Z_2, (1-Z_1)(1-Z_2)Z_3, \ldots). \tag{3.4}$$

Remark: These definitions are for fixed but arbitrary k. The $\{Z_i\}$ and $\{U_j\}$ that appear in (3.2) and (3.3) should perhaps be written $\{Z_i^{(k)}\}$, $\{U_j^{(k)}\}$ to emphasise that they vary with k. We are only interested in distributional properties, and so will continue to suppress the dependence on k.

One can view $\underset{\sim}{F}^{(k)}$ as a random point in Δ^1; we will use it to construct a sequence $\{\nu_k, k = 0,1, \dots\}$ of probability measures on S as follows. ν_k concentrates on $\{k\} \times \Delta^1$, and for any Borel subset A of Δ

$$\nu_k(\{k\} \times A) = P(\underset{\sim}{F}^{(k)} \in A), \; k = 0,1, \dots \tag{3.5}$$

Our key result, which has close affinities with the work of Griffiths [6] is the following theorem.

<u>Theorem</u>:

(i) There is a discrete time Markov chain $\{\eta_k, k = 0,1,\dots\}$ on S such that η_k has distribution ν_k, and

$$P(\eta_{k-1} = (k-1;x_1,\dots,x_{i-1}, x_{i+1},\dots,x_k,x_i,y_1,y_2,\dots)$$
$$|\eta_k = (k;x_1,\dots,x_k,y_1,y_2,\dots,)) = \frac{\theta}{k(k+\theta-1)},$$

$$P(\eta_{k-1} = (k-1;x_1,\dots,x_i+x_j,\dots x_k,y_1,y_2,\dots)$$
$$|\eta_k = (k;x_1,\dots,x_k,y_1,y_2,\dots)) = \frac{2}{k(k+\theta-1)},$$

for $k = 1,2,\dots$; $1 \leq i < j \leq k$.

(ii) Let $\{D_t\}$ be a pure death process with death rate $k(k+\theta-1)/2$ from state k, starting at infinity, and independent of $\{\eta_k\}$. Then the process $\{M_t, t \geq 0\}$ defined by

$$M_0 = (\infty;0,0,\dots), M_t = \eta_{D_t}, \; t > 0$$

is a Markov process on S.

<u>Proof</u>. (i) We need to check the consistency of the finite dimensional distributions of $\{\eta_k\}$. This follows after some lengthy but straightforward calculations using the structure (3.3) of the ν_k's . To establish (ii), we can use an argument similar to that of [9], p. 244.

The process $\{M_t, t \geq 0\}$ may be thought of as a genealogical representation of the infinitely-many neutral alleles diffusion model;

cf. Ethier and Kurtz [4]. Our explicit recognition of the age-ordering of novel alleles leads to a variety of interesting results which we will exploit in the final expository section.

4. APPLICATIONS

(a) Limit distributions

As $t \to \infty$, $D_t \to 0$ a.s., and so $M_t \Rightarrow \nu_0$, defined by (3.4) and (3.5). Eventually, then, the population comprises only new lines of descent, and it follows that at stationarity the frequencies of the oldest, next oldest ,..., alleles in the infinitely-many neutral alleles diffusion model have the representation Z_1, $(1-Z_1)Z_2$, $(1-Z_1)(1-Z_2)Z_3,\ldots$ where the Z_i's are independent and identically distributed r.v.'s with density (3.2). The decreasing order statistics of such a random vector have the Poisson-Dirichlet distribution with parameter θ (cf. [7] and [10]), which distribution is well-known to be the stationary measure of the infinitely-many neutral alleles diffusion model (cf. [4]). The Poisson-Dirichlet distribution is an intractable object to handle explicitly; as well as giving the age ordering of the population frequencies, our approach mitigates some of these difficulties. Donnelly [2] gives a number of interconnections between the distribution ν_0, and the properties of samples taken from such a distribution, exploiting more fully the consequences of size-biased sampling.

(b) K-allele models

Many of the fundamental results about the infinitely-many neutral alleles process were discovered by taking suitable limits in a K-allele model with symmetric mutation, as $K \to \infty$. Here we observe that the fundamental structure of K-allele models with scaled mutation rates $\varepsilon_{ij} \equiv \varepsilon_j$ from alleles of type i to type j (cf. Griffiths [5]) can be recovered from the frequency process $\{M_t, t > 0\}$. The idea is to construct the K allele model from a realisation of $\{M_t, t>0\}$, as in [1]. Lines of descent are initiated by mutations, but now the (age-ordered) classes do not have distinct allelic types. Define $\theta = \varepsilon_1 + \ldots + \varepsilon_K$, and $p_j = \varepsilon_j/\theta$, $j = 1,2,\ldots,K$. Fix $t > 0$, and suppose that the number of old classes, D_t, of M_t is equal to $k > 0$. The relative frequencies of the new classes, normalised to have sum 1, have distribution ν_0 (cf. (3.3) and

(3.4)). Allelic labels $1,2,\ldots,K$ are given to each class frequency by labelling each independently, assigning label j with probability p_j, $j = 1,\ldots,K$. Recall from [7] and [10] that the points Z_1, $(1-Z_1)Z_2,\ldots$ in (3.4) are an enumeration of the points of a non-homogeneous Poisson process with mean measure density $\theta e^{-x}/x$, normalised to have sum 1. It follows that the collections of points $\mathcal{P}_j$, say, of frequencies which are labelled j may be viewed as the (normalised) points of K independent non-homogeneous Poisson processes with mean measure density $p_j \theta e^{-x}/x = \varepsilon_j e^{-x}/x$, $j = 1,\ldots,K$. It can be shown that the total frequencies of new classes labelled j have jointly a K-dimensional Dirichlet distribution, with parameters $\varepsilon_1,\ldots,\varepsilon_K$. Thus when $D_t = k$, the frequencies $X_1,\ldots,X_k$ of old classes, and the frequencies $Y_1,Y_2,\ldots,Y_K$ of alleles of types $1,\ldots,K$ that arose by mutations in $(0,t)$ have the structure

$$(X_1,\ldots,X_k, Y_1,\ldots,Y_K) \overset{d}{=} (U_1V_k,\ldots,U_kV_k,(1-V_k)D_1,\ldots,(1-V_k)D_K), \tag{4.1}$$

where $(D_1,\ldots,D_K)$ have the Dirichlet distribution with parameters $\varepsilon_1,\ldots,\varepsilon_K$, independent of $(U_1,\ldots,U_k)$, and V_k (cf. (3.1)). The structure exhibited by (4.1) was found by different means by Griffiths [5] and it leads immediately to an explicit representation for the transition density of the K-allele diffusion process.

(c) A Population Coalescent

The process $\{M_t,\ t > 0\}$ is closely related to an age-ordered coalescent via a modification of Kingman's paint box scheme. Fix $t > 0$, and assume that $D_t = k$, and $\mathcal{M}_k = (k;x_1,x_2,\ldots)$. Define a sequence of independent and identically distributed random variables $\tau_1,\tau_2,\ldots,$ with

$$P(\tau_1 = r) = x_r, r = 1,2,\ldots\ . \tag{4.2}$$

We can then define a labelled equivalence relation $\mathcal{R}$ on $\mathbb{N}$ as follows. We say that genes i and j are equivalent, in an equivalence class labelled 0, if $\tau_i = \tau_j < k$, while genes i and j are in the equivalence class labelled r if $\xi_i = \xi_j = k + r, r = 1,2,\ldots$. There are thus k equivalence classes labelled 0 (corresponding to the old classes in the introduction) and age-ordered new equivalence classes in which a class with label r is older than a class with label s if $r < s$. Note that by the structure of the measure ν_k, each class is a.s. infinite.

We conclude with a brief discussion of how this process is related to the sample coalescent of section 1. One obtains from $\mathcal{R}$ an equivalence relation in $\mathcal{E}_n^*$ by restricting attention to the labels assigned to genes 1,2 ,..., n. This relation, $\mathcal{Q}^n$ say, will have some number $m(<k)$ of old equivalence classes, and a number ℓ (>0) of new equivalence classes, age ordered in the natural way. It can then be shown that

$$\begin{aligned} P(\mathcal{Q}^n = (\xi_1,\ldots,\xi_m;\eta_1,\ldots,\eta_\ell) \mid D_t = k) \\ = \frac{k!\,\Gamma(k+\theta)\theta^\ell \lambda_1!\ldots\lambda_m!\,\mu_1!\ldots\mu_\ell!}{(k-m)!\,\Gamma(n+k+\theta)\mu_\ell(\mu_\ell+\mu_{\ell-1})\ldots(\mu_\ell+\ldots+\mu_1)} \end{aligned} \tag{4.3}$$

where λ_i is the number of genes in class ξ_i, and μ_i the number of genes in class η_i This result should be compared to that in (2.2), the difference arising from the fact that the number of old classes m in $\mathcal{Q}^n$ may be less than the number of old classes k in $\mathcal{R}$.

ACNOWLEDGEMENT

The authors were supported in part by National Science Foundation grants DMS 85-01763 and DMS 86-08857.

REFERENCES

[1] Donnelly, P. J.(1986) Dual processes in population genetics. In Stochastic Spatial Processes. Ed. P. Tautu. Springer Lecture Notes in Mathematics, in press.

[2] Donnelly, P. J.(1986) Partition structures, Polya urns, the Ewens sampling formula and the ages of alleles. Theor. Popn. Biol., in press.

[3] Donnelly, P. J., Tavare S.(1986) The ages of alleles and a coalescent. Adv. Appl. Prob., 18, 1-19.

[4] Ethier S. N., Kurtz, T.G.(1981) The infinitely-many-neutral-alleles diffusion model. Adv. Appl. Prob. 13, 439-452.

[5] Griffiths, R. C.(1979) A transition density expansion for a multiallele diffusion model. Adv. Appl. Prob., 11, 310-325.

[6] Griffiths, R. C.(1980) Lines of descent in the diffusion approximation of neutral Wright-Fisher models. Theor. Popn. Biol., 17, 37-50.

[7] Kingman, J. F. C.(1975) Random discrete distributions, J. Roy. Statist. Soc., B, 37, 1-22.

[8] Kingman J. F. C.(1982) On the genealogy of large populations. J. Appl. Prob.,19A, 27-43.

[9] Kingman, J. F. C.(1982) The coalescent. Stoch. Proc. Appln., 13, 235-248.

[10] Patil, G. P., Taillie, C.(1977) Diversity as a concept and its implications for random communities. Bull. Internat. Stat. Inst., 47, 497-515.

[11] Watterson, G. A.(1984) Lines of descent and a coalescent. Theor. Popn. Biol., 26, 77-92.