THE GENEALOGY OF THE INFINITE ALLELES MODEL

Peter Donnelly
Department of Statistical Science
University College
Gower Street
London WC1E 6BT, U.K.

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

INTRODUCTION

Our purpose here is to describe a particular stochastic process associated with the genealogy of a (hypothetically infinite) population evolving according to the infinite alleles model of population genetics (on the so called diffusion time scale). The process represents a generalization of both Kingman's (1982b) coalescent, in accounting for mutation, and the n-coalescent with ages of Donnelly and Tavaré (1986) (see also Kingman 1982a and Watterson 1984) to the infinite case. We content ourselves with an outline of the dynamics and the distribution of the process. For a more detailed exposition and several applications see Donnelly and Tavaré (1987). For more of the genetic background and in particular a description of the infinite alleles model see for example Ewens (1979). Our process is associated with any of the models within the domain of attraction of the usual (Wright Fisher) diffusion approximation (see for example Griffiths 1980, to which our work is closely allied). Beyond this we make no specific assumptions about the nature of the reproductive mechanism.

THE PROCESS

Fix $t > 0$ and consider the genealogy of the population at the present time (time 0) with respect to the ancestral population at time $-t$. A certain proportion of the population, x_1 say, may share a common ancestor, without intervening mutation, at time $-t$. A further proportion x_2 say, may share a different common ancestor, again without intervening mutation between time 0 and time $-t$, and so on. Suppose there are k such groups (for $t > 0$, k will be a.s. finite) with frequencies $x_1,\ldots,x_k$. Each such group will be of the same genetic type as the corresponding ancestor. There will also be a number of types in the current population which do not appear in the ancestral population. Each new type will correspond to individuals who share a common ancestor, without intervening mutation, at some time $-r$, $0 < r < t$, that ancestor being a mutant. By assumption each new type will arise at a different time, so order (and label) the new types by age (i.e. time since first appearance), and denote the frequencies of the oldest, second oldest, ..., new type by $x_{k+1},x_{k+2},\ldots$. (For $t > 0$ and non zero mutation rate, there will be an infinite number of new types). For this value of t, describe the above genealogy by the point $(k;x_1,x_2,\ldots,x_k,x_{k+1},x_{k+2},\ldots)$ of $N_\infty \times \Delta$ (where $N_\infty = Nu\{0,\infty\}$ and $\Delta = \{(x_1,x_2,\ldots) : 0 \leqslant x_i \leqslant 1, \sum_{i=1}^{\infty} x_i \leqslant 1\})$. As t varies this gives rise to a stochastic process $\{M_t;t \geqslant 0\}$, and we put $M_0 = (\infty,0,0,\ldots)$.

The process $\{M_t\}$ has a special structure and particularly simple dynamics. It inherits these from the finite n-coalescents with ages of Donnelly and Tavaré (1986) which (as in Kingman 1982b) may be embedded (in two sorts of ways) in $\{M_t\}$. Denote by D_t the number of alleles in common between the population at time 0 and the ancestral population at time $-t$ (so that D_t is the value of k above). Then the process $\{D_t;t \geqslant 0\}$ is a death process with entrance boundary at ∞ and death rates $k(k + \theta - 1)/2$ from state k. (As usual θ denotes the scaled mutation rate.) Furthermore we may construct $\{M_t\}$ by putting $M_0 = (\infty;0,0,\ldots)$ and $M_t = \mathcal{M}_{D_t}$ for

$t > 0$, where $\{\mathcal{M}_k, k = 0,1,\ldots\}$ is a discrete time Markov chain on $N_\infty \times \Delta$, independent of $\{D_t; t \geq 0\}$ with

$$P(\mathcal{M}_{k-1} = (k-1; x_1,\ldots,x_{i-1},x_i + x_j,\ldots,x_{j-1},x_{j+1},\ldots,x_k,x_{k+1},\ldots) |$$
$$\mathcal{M}_k = (k; x_1,\ldots,x_k,x_{k+1},\ldots)) = 2/k(k+\theta-1) , \quad 1 \leq i < j \leq k ,$$

$$P(\mathcal{M}_{k-1} = (k-1; x_1,\ldots,x_{i-1},x_{i+1},\ldots,x_k,x_i,x_{k+1},x_{k+2},\ldots) |$$
$$\mathcal{M}_k = (k; x_1,\ldots,x_k,x_{k+1},\ldots)) = \theta/k(k+\theta-1) , \quad 1 \leq i \leq k .$$

Thus the times between changes of $\{M_t; t \geq 0\}$ depend only on the number of "old" types, and the changes take one of two simple forms: a "coalescence" of two old types or one of the old types becoming "new". The existence of the Markov chain $\{\mathcal{M}_k, k = 0,1,\ldots\}$ is established in Donnelly and Tavaré (1987).

The distribution of $\{D_t; t \geq 0\}$ is well known. See for example Tavaré (1984, equation (5.5)). The distribution, ν_k say, of $\mathcal{M}_k$ is concentrated on $\{k\} \times \Delta^1$ (where $\Delta^1 = \{(x_1,x_2,\ldots) : 0 \leq x_i \leq 1 , \sum_{i=1}^{\infty} x_i = 1\} \subseteq \Delta$) and has the following representation. Let V_k be a random variable having the beta density f_k given by

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

Let $(U_1,U_2,\ldots,U_k)$ be a random k-vector having a uniform distribution on the simplex $\{(u_1,u_2,\ldots,u_k); u_i \geq 0, u_1 + u_2 + \ldots + u_k = 1\}$ and let $Z_1,Z_2,\ldots$ be independent and identically distributed random variables with density f given by

$$f(x) = \theta(1-x)^{\theta-1}, \quad 0 \leq x \leq 1 ,$$

and choose $(U_1,U_2,\ldots,U_k)$, V_k , and $\{Z_j\}$ to be mutually independent. Finally define the random element $X^{(k)}$ of Δ^1 by

$$X^{(k)} = (V_kU_1, V_kU_2, \ldots, V_kU_k, (1-V_k)Z_1, (1-V_k)(1-Z_1)Z_2,$$
$$(1-V_k)(1-Z_1)(1-Z_2)Z_3, \ldots) .$$

When $k = 0$, put

$$X^{(0)} = (Z_1, (1-Z_1)Z_2, (1-Z_1)(1-Z_2)Z_3, \ldots) .$$

Now for any Borel subset A of Δ , define ν_k by

$$\nu_k(\{k\} \times A) = P(X^{(k)} \in A) , \quad k = 0,1,2,\ldots .$$

This form lends itself readily to calculation and is exploited in Donnelly and Tavaré (1987), to recover the transition density of the usual K-allele diffusion models.

ACKNOWLEDGEMENTS

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

REFERENCES

Donnelly, P., Tavaré, S. (1986), The ages of alleles and a coalescent, *Adv. Appl. Prob.* **18**, 1-19.

Donnelly, P., Tavaré, S. (1987), The population genealogy of the infinitely many neutral alleles model. To appear.

Ewens, W.J. (1979), *Mathematical Population Genetics*, Springer, Berlin.

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

Kingman, J.F.C. (1982a), On the genealogy of large populations, *J. Appl. Probl* **18**, 27-43.

Kingman, J.F.C. (1982b), The Coalescent. *Stoch. Proc. Appln.* **13**, 235-248.

Tavaré, S. (1984), Line of descent and genealogical processes and their application in population genetic models. *Theor. Popn. Biol.* **26**, 119-164.

Watterson, G.A. (1984), Lines of descent and the coalescent, *Theor. Popn. Biol.* **26**, 77-92.