AN URN MODEL AND THE COALESCENT IN NEUTRAL INFINITE-ALLELES GENETIC PROCESSES

David Branson
University of Essex

Abstract

An urn model is proposed as a useful computational device for obtaining results within the context of infinite-alleles genetic processes incorporating selective neutrality. The model is based on Kingman's coalescent, as developed by Watterson and by Donnelly and Tavaré, and mimics the procedure of tracing a sample's ancestry backwards in time, noting the appearance of common ancestors or new mutants. The structure of the model is possessed by a time-inhomogeneous linear birth-and-death model with immigration, by Moran's model and by a class of other models including the Wright-Fisher model in the limit of large population size. Elementary combinatorial arguments connected with the partial or complete emptying of the urn give rise to a number of results which can be interpreted in terms of the allelic composition of genetic samples and populations. The relationship to other urn models is discussed.

1. Introduction. Some years ago Kingman (1982 a,b,c) introduced the fruitful concept of the coalescent in which a sample of genes is taken from a population and its ancestry traced back in time, noting where there are common ancestors, until one reaches a single common ancestor. This idea has been developed by others, in particular by Watterson (1984) and by Donnelly and Tavaré (1986) who consider in detail the consequences of mutation. In tracing the ancestry backwards we may come to a mutant which introduced a new allele into the population and which was therefore the originating ancestor for a family represented in the sample, all members of the family carrying the same allele. We assume mutation is non-recurrent, that is each mutation produces a previously unknown allele and we assume that selection is absent. We further suppose that the organism we are considering is a monoecious haploid reproducing either asexually or by self-fertilization, so that each individual has just one parent and, in the absence of mutation, the genetic composition of an individual is identical to that of its parent.

In Section 2 we describe an urn model which mimics the coalescent and which keeps track of the time-ordering of the various mutant and non-mutant births, should this be of interest. The model gives a method of constructing the jump chain of Donnelly and Tavaré's coalescent with ages and is put forward as

a way of visualizing the process and as a useful computational device for deriving results (particularly when the urn is only partly emptied) which can be interpreted in terms of a number of population models. This is illustrated in Section 4 where elementary combinatorial arguments provide a simple derivation of several results most of which have appeared elsewhere in various contexts.

Donnelly and Tavaré show that their model (and hence our urn model) describes Moran's model and a class of other models with non-overlapping generations including the Wright-Fisher model in the limit of large population size. Tavaré (1989) has shown that the coalescent with ages reproduces the results of the (time-homogeneous) birth-death-immigration model conditioned on population size. In this case the possibility of mutation within a family is excluded, but the introduction of a new allele into the population is modelled by the immigration of an individual carrying this allele, who then becomes the originating ancestor of a family. We show in Section 3 that the structure of our urn model is also possessed by a time-inhomogeneous birth-death-immigration model provided the birth-per-individual and immigration rates are proportional; we discuss the validity of this assumption. For the Moran and birth-death-immigration models, both the population as a whole and a sample chosen from it have the structure of our urn model.

In Section 5 a brief discussion is given of other urn models in the literature of population genetics. We place our model in the context of these others, and point out its differences from them.

2. An Urn Model. The coalescent takes (without replacement) a sample from a population and traces their ancestry back in time, noting where there are common ancestors. We allow the possibility of mutation (or immigration, but for convenience we shall use the word mutation to refer to either sort of event), so in tracing ancestry backwards we may come to a mutant who was the originating ancestor for a family represented in the sample; we trace this branch of the population back no further. That is, we are not concerned here with the prior history of an individual who introduced a previously unknown allele into the population.

We call the pattern obtained by tracing the sample back through the generations to the original ancestor(s) the sample-forest, as it will consist of a collection of trees each headed by a new mutant. We take no notice of a non-mutant birth with the simultaneous death of the parent (such as is allowed in the Moran model) since this leads neither to a node in the sample-forest, nor to the termination of a tree - in this case we simply identify parent and offspring; neither are

we interested here in anything outside the sample-forest. In fact in tracing the sample-forest backwards in time we take account of only two categories of event:

(i) coalescence: two branches in the sample-forest merge at a birth (with no mutation) either because two individuals belonging to the forest at a particular time are a parent-offspring pair, or because, in a model where individuals do not survive the birth of their offspring, two members of the sample-forest in one generation have the same parent in the previous generation. In order to unify our treatment we find it helpful in the latter case to pretend that one of the two individuals in the later generation is a surviving parent and the other is its offspring. This is clearly a harmless fiction since parent and offspring are genetically identical.

(ii) mutation: one branch in the sample-forest ends with the birth of a mutant.

In Figure 1 the thick lines show the sample-forest arising from the six individuals denoted by black squares. The sample-forest contains three coalescences and three mutations; it consists of three trees (families) containing one, two and three members. The members of each family are genetically identical, but the families are genetically distinct.

The probability of simultaneous events of type (i) or (ii) is taken to be zero and all events are assumed to be independent. We shall often refer to "survivors" at a particular time: by this we mean the branches of the sample-forest which, in tracing backwards to the time in question, have not been removed by coalescence or mutation. Each event reduces by one the number of survivors, so that, if the original sample was of size m, the forest ends after m events.

We now describe an urn model whose relevance to these ideas will be detailed in the following section. Suppose an urn originally contains m balls of the same size but each of a different colour (not black) representing a sample of size m taken from a population. We remove the balls from the urn in m stages, which we label successively as stage m, stage $m-1$, ..., stage 1 (so that stage i commences with i balls in the urn). In stage m we remove one ball chosen at random and designate it an "offspring" ball. We replace it with a black ball whose size is such that the relative probabilities of picking the black ball or any particular coloured ball are in the ratio $\theta:1$. We now remove a second ball: if it is black, stage m ends (the interpretation will be that the "offspring" was a mutant); if it is coloured, we designate it a "parent" ball, replace it in the urn, remove the black ball, but recolour the "offspring" ball with the colour of its "parent" (this will be

interpreted as a coalescence as described above). This completes stage m.

We continue in a similar fashion. In stage i, the urn contains i balls, and the probability that any particular ball is chosen as "offspring", followed by the black ball is

$$\frac{1}{i}\frac{\theta}{\theta+i-1} \quad . \tag{2.1}$$

On the other hand, the probability that any particular ball is chosen as "offspring" and any other particular ball as "parent" is

$$\frac{1}{i}\frac{1}{\theta+i-1}. \tag{2.2}$$

If any stage results in a coalescence we recolour with the "parent's" colour not only the "offspring" at that stage but also all previously removed balls of the same colour (which are "offspring" of "offspring" ...). For example, at stage m we may remove the blue ball as offspring and the red ball as parent, so we replace the red ball but change the colour of the offspring ball from blue to red. At a later stage we may remove the red ball as offspring and the green ball as parent. We replace the green ball but change to green the colour of both balls that are now red (that is not only green's "child" but also its "grandchild" originally coloured blue). Thus at any stage we have painted with the same colour a parent (in the urn) and all its progeny of succeeding generations (outside the urn).

After stage 1 the urn will be empty. Outside the urn will be m balls of l (say) different colours if the black ball was chosen on l occasions. The balls carrying the same colour represent a single family descended from an originating mutant ancestor. That is, in genetic terms, the sample contains l distinct alleles.

3. Population Models. The urn model described in Section 2 will apply to any population model whose transitions from one stage to the next are independent with probabilities given by (2.1) and (2.2). Donnelly and Tavaré (1986), in discussing their coalescent with ages, show that this is the case for Moran's model if we put

$$\theta = Mu/(1-u), \tag{3.1}$$

where M is the population size and u is the probability that any offspring is a mutant. They show it is also the case for a class of models with non-overlapping generations including the large population limit of the Wright-Fisher model for which one lets M tend to infinity and u tend to zero with θ=2Mu finite and non-zero. (Note that their expressions equivalent to (2.2) contain a factor 2 as, fol-

lowing a birth, they do not distinguish between offspring and parent.) Tavare (1989) has discussed the connection between the coalescent with ages and a time-homogeneous birth-death-immigration model where each immigrant is the source of a new allele in the population. However it is likely that fertility, mortality and immigration will vary in time in response to (perhaps seasonal) variations in the environment. Let us therefore suppose that the birth-per-individual, death-per-individual and immigration rates at time t are given respectively by $\lambda(t)$, $\mu(t)$ and $\nu(t)$. The functions $\lambda(t)$ and $\mu(t)$ are arbitrary functions of time, but the same functions for different families; that is, the different alleles possessed by different families confer no advantage or disadvantage concerning birth or death rates. In other words, our model incorporates selective neutrality. For convenience, after a birth has occurred, we arbitrarily label one of the two resulting identical individuals as "parent" and the other as "offspring". Denote the population size at time t by M(t).

We take (without replacement) a sample of size m, and trace its ancestry back in time, as described in Section 2, taking note of coalescences and mutations (immigrations). Suppose stage i+1 has been completed, leaving i survivors, and suppose that a birth or immigration in the population occurs at time τ, decreasing the population size from $M(\tau)$ to $M(\tau)-1$. The probabilities of this event's being an immigration or a birth are, respectively,

$$\nu(\tau)/[\nu(\tau)+(M(\tau)-1)\lambda(\tau)], \qquad (M(\tau)-1)\lambda(\tau)/[\nu(\tau)+(M(\tau)-1)\lambda(\tau)].$$

Since each of the M individuals in the population is equally likely to be the new arrival, and each individual is equally likely to be the parent it follows that the probability that any particular individual is a new immigrant is

$$\frac{\nu}{M[\nu+(M-1)\lambda]} \tag{3.2}$$

whereas the probability that any two particular individuals are respectively a new-born individual and its parent is

$$\frac{(M-1)\lambda}{M(M-1)[\nu+(M-1)\lambda]}. \tag{3.3}$$

From (3.2) and (3.3) we see that the probability that the individuals involved in the event are taken from the i current survivors is

$$\frac{i\nu+i(i-1)\lambda}{M[\nu+(M-1)\lambda]}. \tag{3.4}$$

Conditioning on the event whose probability is given in (3.4) we conclude from

(3.2) and (3.3) that, if stage i occurs at time τ, the probability that it involves the immigration of any particular survivor is

$$\frac{1}{i}\,\frac{\nu}{\nu+(i-1)\lambda} \tag{3.5}$$

whereas the probability that it involves any particular survivor as offspring and any other as its parent is

$$\frac{1}{i}\,\frac{\lambda}{\nu+(i-1)\lambda}. \tag{3.6}$$

If $\nu(\tau) = \theta\lambda(\tau)$ for some constant θ, the probabilities in (3.5) and (3.6) become independent of time and are given respectively by

$$\frac{1}{i}\,\frac{\theta}{\theta+i-1} \tag{3.7}$$

and

$$\frac{1}{i}\,\frac{1}{\theta+i-1}\ . \tag{3.8}$$

Fundamental to our argument above is the Markovian nature of the process (which also guarantees the independence of the different stages in the sample-forest) and our assumption of selective neutrality, which ensures that all individuals have the same fertility and mortality. This is our justification for supposing that all possible parents of an individual are equally probable. Note that (3.7) and (3.8) depend only on the ratio of the immigration and birth rates and are quite independent of the death rate. Also our discussion does not require the establishment of an equilibrium. We see that (3.7) and (3.8) are identical to (2.1) and (2.2), so the urn model of Section 2 can be applied to this time-dependent birth-death-immigration model.

The requirement that $\nu(t)=\theta\lambda(t)$ is not as artificial as it may seem. Generally, a more favorable environment leading to a higher birth rate may well attract a higher rate of immigration, or both rates may undergo the same periodic variation. More quantitatively, we may suppose that the "immigrants" are in fact mutant offspring. Denote by $\Lambda(t)$ the birth rate (per head) of any offspring, whether mutant or not. Suppose that any offspring, independent of any other, has probability $u(t)$ of being a (previously unknown) mutant. Then $\lambda(t)$, the birth rate of non-mutant offspring is given by

$$\lambda(t) = [1-u(t)]\Lambda(t)$$

and $\nu(t)$, the rate of appearance in the population of "immigrants" (that is, mutant offspring) is given by

$$\nu(t) = M(t)u(t)\Lambda(t).$$

Hence $\nu(t)=\theta\lambda(t)$ provided the probability of mutation is related to the population size by the same relation (3.1) as in Moran's model. This is the same sort of relationship implicit in diffusion models. The validity of such assumptions has been discussed by Karlin and McGregor (1964), Crow (1985) and Gillespie (1989).

Originally we took a sample of size m, but there is no reason why m may not be equal to the total population size M. This is also the case for Moran's model. Thus, for these models, any results derived apply not only to a sample, but also to the population as a whole.

4. Distributions of Genetic Properties. A number of authors have obtained results concerning genetic distributions in the context of various models. The purpose of this section is to show how the urn model provides via straightforward combinatorics a simple computational tool for deriving such results for any genetic model compatible with (2.1) and (2.2). Most of the results we discuss have been obtained elsewhere by other methods, but Theorem 4.12 is believed to be new. We shall refer to equations and other results in Donnelly and Tavaré (1986), Kelly (1977) and Watterson (1984) by the prefixes DT, K and W respectively.

We take a sample of size m, without replacement, and trace its ancestry back in time as previously discussed. (We again emphasize that in the birth-death-immigration and Moran models m may equal the total population size.) However for most of our results we now terminate the procedure after stage k+1 (that is, after $m-k$ stages) at a time when there are k survivors each of whom heads a genetically identical family consisting of a particular subset of the original sample. In Watterson's (1984) terminology these are "old" genes. Let $\lambda_1, \lambda_2, \ldots, \lambda_k$ be the sizes of these families. By this stage a number, l (say), of mutants have appeared each heading a genetically identical family of size $\mu_1, \mu_2, \ldots, \mu_l$. These are Watterson's "new" genes. Let z be the total number of old genes, so that

$$z = \Sigma\lambda_i = m - \Sigma\mu_i. \tag{4.1}$$

THEOREM 4.1 (DT 2.8). We label in some way the genes in the original sample of size m. The probability that, after stage k+1, the old gene families consist of specified genes and are of sizes $\lambda_1, \lambda_2, \ldots, \lambda_k$ and that there are l families of

new genes (again consisting of specified individuals) of sizes $\mu_1, \mu_2, \ldots, \mu_l$ *in order of increasing allele age* (i.e. the mutant originator of the family of size μ_1 is younger than the mutant originator of the family of size μ_2, and so on) is

$$\frac{(m-k)!\ \ k!\ \ \theta^l}{m!\ (\theta+k)^{[m-k]}}\ \frac{\lambda_1!\lambda_2!\ldots\lambda_k!\mu_1!\mu_2!\ldots\mu_l!}{(\mu_1+\ldots+\mu_l)\ldots(\mu_1+\mu_2)\mu_1}\quad , \tag{4.2}$$

where θ has the value appropriate to the particular model under consideration, as described in Section 3, and the rising factorial notation is defined by

$$(x)^{[n]} = x(x+1)\ (x+2)\ldots(x+\ n-1).$$

We give a proof of this result via the urn model. For stage i, commencing with i balls in the urn, the probabilities that any particular ball is chosen as offspring followed by respectively the black ball or any particular parent ball are given by (2.1) and (2.2). Following the original sample through stages m, m − 1,...,k+1 corresponds to operating the urn model until k balls remain in the urn, and the appearance of l mutants corresponds to the black ball's having been chosen on l occasions. Therefore, from (2.1) and (2.2), we see that the probability of any particular sequence leading to this outcome is

$$\frac{1}{m(m-1)\ldots(k+1)}\ \frac{\theta^l}{(\theta+m-1)\ (\theta+m-2)\ldots(\theta+k)}$$

$$= \frac{k!\,\theta^l}{m!\ (\theta+k)^{[m-k]}}. \tag{4.3}$$

In order to calculate the number of (equally probable) paths giving the result specified in the statement of the theorem, we first calculate the number of orders in which we can choose the offspring balls. We designate balls as "old" or "new" according as the corresponding genes are "old" or "new". We choose one ball from each of the old families (in $\lambda_1\ \lambda_2\ \ldots\ \lambda_k$ ways). These will be the k balls finally left in the urn. We number the remaining balls 1, 2, ..., m–k as described below and withdraw the offspring balls in this order. We first number the remaining old balls at random (in (m–k)!/(m–z)! ways). We next choose one ball from each of the new families (in $\mu_1\ \mu_2\ \ldots\ \mu_l$ ways). These will be the oldest members of each new family. We now number the new balls with the remaining m–z numbers starting with the oldest family (of size μ_l) and finishing with the youngest family (of size μ_1). The allocation of numbers here is random except that when we come to a fresh family we first give its oldest member the highest

available number. This clearly guarantees the required age structure of the new families, and can be done in $(m - z)!/[(\mu_1 + ... + \mu_l)\ (\mu_1 + ... + \mu_{l-1})\ ...\ (\mu_1 + \mu_2)\mu_1]$ ways.

Within any family the parent balls are successively chosen from a diminishing pool of $(\lambda-1),(\lambda-2)$, ... (or $(\mu-1)$, $(\mu-2)$, ...) potential parents, so the parent balls can be chosen in $(\lambda_1-1)!...(\lambda_k-1)!(\mu_1-1)!...(\mu_l-1)!$ ways. Putting all these factors together, we find that the number of ways we can obtain the desired outcome is

$$\frac{(m-k)!\,\lambda_1!\,\lambda_2!\,...\,\lambda_k!\,\mu_1!\,\mu_2!\,...\,\mu_l!}{(\mu_1 + ... + \mu_l)\,...\,(\mu_1 + \mu_2)\,\mu_1}. \tag{4.4}$$

If we multiply (4.4) by the probability of any path (4.3) we obtain the result (4.2).

THEOREM 4.2 (W 2.9). The situation is the same as in Theorem 4.1 except that ages of the new families are now immaterial. The probability is now

$$\frac{(m-k)!\,k!\,\theta^l}{m!\,(\theta + k)^{[m-k]}}\ \lambda_1!\,\lambda_2!\,...\,\lambda_k!\,(\mu_1 - 1)!\,(\mu_2 - 1)!\,...\,(\mu_l - 1)!\,. \tag{4.5}$$

The proof is the same as in Theorem 4.1 except for the numbering of the balls. Once we have chosen the k (old) balls to be left in the urn, we simply number at random all the remaining m-k (old and new) balls. This can be done in (m – k)! ways and the result follows immediately.

THEOREM 4.3 (W 3.3.1). As in Theorem 4.2, except that we do not specify the particular genes belonging to any family. We simply require k old families of sizes λ_1, ..., λ_k and l new families of sizes μ_1, ..., μ_l. Denote by α_j (β_j) the number of old (new) families of size j. The probability of this configuration is

$$\frac{(m-k)!\,k!\,\theta^l}{(\theta + k)^{[m-k]}}\ \frac{1}{\mu_1\mu_2...\mu_l\ \alpha_1!\,\alpha_2!\,...\,\alpha_m!\ \ \beta_1!\,\beta_2!\,...\,\beta_m!}. \tag{4.6}$$

To prove this, we note that the number of ways m balls can be distributed amongst families of the given sizes is

$$m!/[\lambda_1!\,\lambda_2!\,...\,\lambda_k!\mu_1!\mu_2!\,...\,\mu_l!\alpha_1!\alpha_2!\,...\,\alpha_m!\,\beta_1!\beta_2!...\beta_m!]. \tag{4.7}$$

The $\lambda_i!$ or $\mu_i!$ factors account for the fact that the ordering within a family is immaterial and the $\alpha_j!$ or $\beta_j!$ factors account for the fact that the ordering of families of the same type containing the same number of members is immaterial. If

we multiply (4.5) by (4.7) we obtain (4.6).

Similarly we can multiply (4.2) by $m!/[\lambda_1!\lambda_2!...\lambda_k!\mu_1!\mu_2!...\mu_l!\alpha_1!\alpha_2!...\alpha_m!]$ to obtain a corresponding result where the allocation of particular genes to families is unspecified but the age ordering of the new families is retained.

THEOREM 4.4 (W 3.4.1). The probability that, after stage k+1 (when k old families remain), there are α_j old families of size j (j=1, 2, ..., m) is given by

$$\Pr(\alpha_1,\alpha_2, ..., \alpha_m \mid k) = \frac{(m-k)!\,k!}{(\theta+k)^{[m-k]}} \binom{\theta+m-z-1}{m-z} \frac{1}{\alpha_1!\,\alpha_2!\ldots\alpha_m!} \quad (4.8)$$

where $z = \Sigma j\alpha_j$ and $k = \Sigma\alpha_j$.

First, for a particular division into old and new balls, we obtain the probability of a particular sequence of offspring balls and old parent balls, summed over all possibilities for new parent balls. The probability of picking any particular sequence of m–k offspring balls is $[m(m-1)...(k+1)]^{-1} = k!/m!$. Suppose that after i stages we have removed j balls destined to be new so that the urn contains m–i balls of which $m-z-j$ are destined to be new. After the next offspring ball has been chosen, there will be $m-i-1$ balls in the urn, to which is added the black ball. A second ball is now withdrawn. If the offspring ball was destined to be old, the probability of picking any particular parent ball destined to be old is $(\theta + m - i - 1)^{-1}$. On the other hand, if the offspring ball was destined to be new, the second ball picked from the urn may be the black ball or any of the $m-z-j-1$ remaining balls that are potentially new parents. The sum of the corresponding probabilities is $(\theta+m-z-j-1)\,/\,(\theta+m-i-1)$. Bearing in mind that there are a total of $m-k$ stages of which m–z involve new balls, we see that the probability of a particular sequence of old parent balls summed over all possibilities for new parents is

$$\frac{(\theta+m-z-1)\,(\theta+m-z-2)\ldots\theta}{(\theta+m-1)\,(\theta+m-2)\ldots(\theta+k)} = \frac{(m-z)!}{(\theta+k)^{[m-k]}} \binom{\theta+m-z-1}{m-z}. \quad (4.9)$$

We must now calculate how many such sequences give rise to an outcome conforming to the statement of the theorem. We can choose the old balls in $\binom{m}{z}$ ways. These balls can be distributed in families of the required sizes, $\lambda_1, \lambda_2, ...,$ λ_k in $z!/(\lambda_1!\lambda_2!...\lambda_k!\alpha_1!\alpha_2!...\alpha_m!)$ ways (compare (4.7)). The k balls which are finally left in the urn, one from each old family, can be chosen in $\lambda_1\lambda_2\ldots\lambda_k$

ways. As before, the offspring balls may be picked in $(m-k)!$ different orders, whereas the old parent balls may be chosen in $(\lambda_1 - 1)!...(\lambda_k - 1)!$ different orders. Assembling all the factors and simplifying gives (4.8).

THEOREM 4.5 (W 3.6.2, but note that Watterson's equation apparently contains an error which we correct here.). The probability that after stage k+1 there are l new families of sizes μ_1, ..., μ_l such that β_j of these families have j members (j=1,2,...,m) is given by

$$\Pr(\beta_1,\beta_2,...,\beta_m \mid k) = \frac{(m-k)!\,\theta^l}{(\theta+k)^{[m-k]}} \binom{z-1}{z-k} \frac{1}{\mu_1\mu_2...\mu_l\beta_1!\,\beta_2!...\beta_m!}, \quad (4.10)$$

where $l = \Sigma\beta_j$ and z is given by (4.1).

The argument parallels that of the previous theorem. First, for a particular division into old and new balls, we find the probability of a particular sequence of offspring balls and new parent balls, summed over all possibilities for old parent balls. The probability of any particular offspring sequence is, as before, k!/m!. Suppose now that after i stages we have removed j balls destined to be *old*, leaving z–j. If the next offspring ball to be chosen is destined to be new, the second ball chosen will be either black, with probability $\theta/(\theta+m-i-1)$, or any particular parent ball destined to be new, with probability $(\theta+m-i-1)^{-1}$. If the offspring ball was destined to be old, the parent ball may be any of the (z–j–1) remaining balls destined to be old. Since the black ball will be chosen l times and a total of z–k old parents will be chosen, the probability of a particular sequence of new parent balls summed over all possibilities for old parents is

$$\frac{(z-1)\,(z-2)...k\quad\theta^l}{(\theta+m-1)\,(\theta+m-2)...\,(\theta+k)} = \frac{(z-1)!}{(k-1)!}\,\frac{\theta^l}{(\theta+k)^{[m-k]}} \quad (4.11)$$

We can choose the old balls in $\binom{m}{z}$ ways, and from these the k old balls finally remaining in the urn can be chosen in $\binom{z}{k}$ ways. The m–z new balls can be distributed in families of the required sizes in $(m-z)!/(\mu_1!\mu_2!...\mu_l!\,\beta_1!\beta_2!...\beta_m!)$ ways. Finally, the offspring balls may be picked in $(m-k)!$ different orders and the new parent balls in $(\mu_1-1)!(\mu_2-1)!\,...(\mu_l-1)!$ different orders. Putting these factors together and simplifying gives (4.10).

THEOREM 4.6 (W 3.4.5) The probability that after stage k+1 there are a total of z old genes (and hence m–z new genes) is given by an inverse hypergeometric distribution:

$$\Pr(z \mid k) = \binom{z-1}{k-1}\binom{\theta+m-z-1}{m-z} \Big/ \binom{\theta+m-1}{m-k}. \tag{4.12}$$

The proof combines arguments from Theorems 4.4 and 4.5. For a particular division into old and new balls, we find the probability of a particular sequence of offspring balls, but now summed over all possibilities for parent balls. The probability of any particular offspring sequence is k!/m!. The arguments leading up to (4.9) and (4.11) show that the sum of all probabilities for the $z-k$ old parents and m–z new parents or black balls is

$$\frac{[(z-1)(z-2)\ldots k]\,[(\theta+m-z-1)(\theta+m-z-2)\ldots\theta]}{(\theta+m-1)(\theta+m-2)\ldots(\theta+k)} \tag{4.13}$$

$$= \frac{(z-1)!\,(m-z)!}{(k-1)!\,(m-k)!} \frac{\binom{\theta+m-z-1}{m-z}}{\binom{\theta+m-1}{m-k}}.$$

As before we can choose the old balls in $\binom{m}{z}$ ways, from these the k old balls finally remaining in the urn can be chosen in $\binom{z}{k}$ ways, and we can pick the offspring balls in $(m-k)!$ different orders. These factors together give (4.12).

THEOREM 4.7 The probability that after stage k+1 there are z old genes and l families of new genes is given by

$$\Pr(z, l \mid k) = \binom{z-1}{k-1}\frac{\theta^l\, c(m-z,\, l)}{(m-z)!} \Big/ \binom{\theta+m-1}{m-k}, \tag{4.14}$$

where $c(m-z, l)$ is a signless Stirling number of the first kind defined by

$$(x)^{[n]} = \sum_{l=0}^{n} c(n, l)\, x^l.$$

The argument is identical to that for the previous theorem except for the choice of new parent balls. Now the black ball is chosen on l occasions and a non-black ball on the remaining (m–z–l) occasions. Hence the factor $(\theta+m-z-1)(\theta+m-z-2)\ldots\theta$ in (4.13) must be replaced by

$$\theta^l \sum a_1 a_2 \ldots a_{m-z-l}$$

where the sum is taken over all a_i satisfying

$$m-z-1 \geq a_1 > a_2 > \ldots > a_{m-z-l} \geq 1$$

and is equal to c(m–z, *l*). This replacement transforms (4.12) into (4.14).

THEOREM 4.8 (W 3.4.6).

$$\Pr(\alpha_1,\alpha_2,\ldots,\alpha_m \mid z,k) = \frac{k!}{\alpha_1!\,\alpha_2!\ldots\alpha_m!} \Big/ \binom{z-1}{k-1}. \tag{4.15}$$

THEOREM 4.9 (W 3.5.1).

$$\Pr(\beta_1,\beta_2,\ldots,\beta_m \mid z, k) = \frac{(m-z)!\,\theta^l}{\theta^{[m-z]}} \quad \frac{1}{\mu_1\mu_2\ldots\mu_l\ \beta_1!\,\beta_2!\ldots\beta_m!}. \tag{4.16}$$

As Watterson points out, this is Ewens' (1972) formula applied to the m-z new genes.

THEOREM 4.10 (W 3.5.2).

$$\Pr(l \mid z,k) = \theta^l c(m-z,l)/\theta^{[m-z]}. \tag{4.17}$$

These three theorems follow immediately from, respectively, Theorems 4.4, 4.5 or 4.7 and Theorem 4.6.

Similarly, Theorems 4.3 and 4.6 show that $\Pr(\alpha_1,\alpha_2,\ldots,\alpha_m,\beta_1,\beta_2,\ldots,\beta_m \mid z,k)$ is given by the product of the right-hand sides of (4.15) and (4.16) and, using (4.17) and (4.16), we can readily obtain an expression for $\Pr(\beta_1,\beta_2,\ldots,\beta_m \mid l,z,k)$.

In any of the foregoing theorems we may of course set k equal to zero to obtain results corresponding to a complete tracing of the sample's ancestry through m stages until no old families remain.

THEOREM 4.11 (DT, Remark, page 14; K, Result 7). We take (without replacement) a sample of size m. (As we have previously emphasized, in appropriate models m may equal the total population size.) Tracing the ancestry of the sample through m stages we discover as usual a number of distinct families each of which was founded some time in the past by the introduction of a previously unknown allele. The probability that the oldest allele in the sample is represented by z members of the sample is

$$(\theta/m)\binom{m}{z} \Big/ \binom{\theta+m-1}{z}.$$

This is found, after a little manipulation, from (4.12) with k set equal to one, since, when only one old family remains, this family necessarily contains the oldest allele in the sample.

THEOREM 4.12 The probability that the youngest allele in the sample is represented by y members of the sample is given by a truncated Yule distribution:

$$\frac{\theta}{y} \Big/ \binom{\theta+y}{y} \qquad y = 1,2,\ldots, m-1 \qquad (4.18)$$

$$\binom{\theta+m-1}{m-1}^{-1} \qquad y = m. \qquad (4.19)$$

Suppose that after stage k+1 there is one new family and it has y members. This could happen either because such a family originated at stage k+1 or because there was one new family with y members at the end of stage k+2, and the parent ball at stage k+1 was non-black (with probability $k/(\theta+k)$). Hence the probability that the youngest family in the sample contains y members and was introduced into the sample-forest at stage k+1 is

$$\Pr(z=m-y, l=1 \mid k) - \Pr(z=m-y, l=1 \mid k+1)[k/(\theta+k)].$$

Using (4.14) and the fact that c(y,1)=(y−1)! this can be written as

$$\left[\frac{y}{\theta+k}\binom{m-y-1}{k-1}\binom{\theta+y}{y}\Big/\binom{\theta+m-1}{m-k-1}\right]\left[\frac{\theta}{y}\Big/\binom{\theta+y}{y}\right]. \qquad (4.20)$$

If y=m, then necessarily k=0 and (4.20) reduces to (4.19). If y<m then to obtain the required probability we must sum (4.20) over the possible values of k=1,2,...,m−y. The terms in the first square bracket in (4.20) are in the form of an inverse hypergeometric distribution in k−1 and therefore sum to one, leaving the second square bracket which is equal to (4.18).

THEOREM 4.13 (DT Proposition 4.1; K, Result 10). Take a sample of size m and trace its ancestry back through m stages, thereby partitioning the sample into genetically distinct families. Suppose one such family contains j members. Then the allele carried by this family is the oldest in the sample with probability j/m.

The urn model provides a trivial proof of this result. The oldest allele in the sample will be that carried by the last ball to be removed from the urn. Each ball has equal probability 1/m of being the last to be removed from the urn. Therefore a family with j members contains the last ball to be removed with

probability j/m.

5. Relationship to Other Urn Models. Ewens (1972) laid the foundation for the use of urn models in population genetics in his treatment of the infinite-alleles Wright-Fisher model at equilibrium. His intuitive argument showed that, in drawing a sample one by one from the population, the probability that the ith gene drawn is of a novel allelic type not seen on the first $i-1$ draws is $\theta/(\theta+i-1)$, independent of the allelic composition of the first $i-1$ genes.

The urn idea was made explicit by Hoppe (1984) and Watterson (1985). The ingredient they added was the specification of the allelic type of the ith gene if it is not novel: they realized that its type is simply uniformly chosen from the genes (not alleles) already present in the sample. Specifically, the probability that the ith gene is of a particular allelic type already existing in the sample is $\mu/(\theta+i-1)$, where μ is the number of genes amongst the first $i-1$ to be drawn that are of that particular allelic type. Thus, in order to simulate a sample of size m, Hoppe considers an urn originally containing a black ball of mass θ. At each of m successive steps a ball is drawn from the urn (with probability proportional to its mass). If the black ball is drawn, it is replaced together with a ball (of unit mass) of a new colour; if a coloured (not black) ball is drawn, it is replaced together with an identical ball. After m steps the allelic composition of the sample is represented by the colour composition of the non-black balls in the urn. Watterson's construction is equivalent.

This model was investigated in considerable detail by Donnelly (1986) and Hoppe (1987). They showed that the urn model mimics the drawing of a sample, at some fixed time, from a Poisson-Dirichlet population, but they also established a connection with the time development of the underlying stochastic process: take a sample of size m and completely trace its coalescent backwards in time through m stages until no old genes remain; Donnelly and Hoppe show that Hoppe's urn model (for a sample of size m) is equivalent to the *reversed* jump chain of the coalescent. The labelling of the colours of the first m balls in his urn model according to the order of their appearance is probabilistically equivalent to labelling the allelic types in the time-reversed coalescent according to their appearance in genealogical time, i.e. according to their relative ages. This correspondence allowed both Donnelly and Hoppe to rederive in a simple fashion a number of known results.

By contrast, the urn model described in Section 2 directly mimics the coalescent itself (rather than its time reversal). If necessary, it can also keep track of the ages of alleles (as, for example, in Theorems 4.1, 4.11, 4.12 and 4.13). If the

coalescent runs its full course, leaving no old genes, (that is, in our notation, k=z=0) then it is often unimportant which way we consider time to run and many results can be obtained equally well by Hoppe's model or by our model. For example results 3, 9 and 11 of Section 5 of Hoppe (1987) are respectively the same as our Theorem 4.13, Theorem 4.10 (with k=z=0, m=2, l=1) and Theorem 4.10 (with k=z=0). However, the new feature of our model is that it provides a simple computational tool in those cases (such as those considered by Watterson (1984) and Donnelly and Tavare (1986)) where, in tracing backwards through time, the coalescent does not run its full course and so there are some old genes remaining. This corresponds, in our model, to the urn being only partly emptied.

In a more recent paper, Watterson (1989) considers the situation where θ varies in time because of a varying population size or mutation rate (or both). There is nothing intrinsic to the coalescent or to our urn model that requires θ to be constant; we simply require all genes, at any given time, to be equally likely to die, give birth or mutate. Therefore we can adapt our model to include the case discussed by Watterson. Suppose θ is piecewise constant, taking values $\theta_1,\theta_2,\theta_3,...$ during successive time periods of duration $t_1,t_2,t_3,...$ proceeding from the present into the past. Let $l_1, l_2, l_3,...$ be the (random) numbers of survivors - in the terminology of Section 2 - at the times when the value of θ changes. Watterson discusses the distribution and simulation of these numbers. Conditional on the values of $l_1, l_2, l_3,...$ we can reproduce Watterson's result (2.2.11) by operating our urn model in the usual fashion, but changing the relative probability of picking the black ball from θ_k to θ_{k+1} after $m - l_k$ stages, where m is the size of the sample. As usual, our model can be terminated before the urn is empty thereby producing results concerning the joint distribution of old and new genes.

Watterson (1987) has discussed yet another urn model which he uses to simulate samples from a population with both neutral and deleterious allelic types. The coalescent, and hence our urn model, are based in an essential fashion on selective neutrality and hence cannot be used in this case.

References

[1] Crow, J. F. (1985). The neutrality-selection controversy in the history of evolution and population genetics. In *Population Genetics and Molecular Evolution* (T. Ohta and K. Aoki, eds.) 1-18. Japan Sci. Soc. Press, Tokyo/ Springer-Verlag, Berlin.

[2] Donnelly, P. (1986). Partition structures, Polya urns, the Ewens sampling formula, and the ages of alleles. *Theoret. Population Biol.* 30 271-288.

[3] Donnelly, P. and Tavaré, S. (1986). The ages of alleles and a coalescent. *Adv. in Appl. Probab.* 18 1-19.

[4] Ewens, W. J. (1972). The sampling theory of selectively neutral alleles. *Theoret. Population Biol.* 3 87-112.

[5] Gillespie, J. H. (1989). When not to use diffusion processes in population genetics. In *Mathematical Evolutionary Theory* (M. W. Feldman, ed.) 57-70. Princeton U.P.

[6] Hoppe, F. M. (1984). Polya-like urns and the Ewens' sampling formula. *J. Math. Biol.* 20 91-94.

[7] Hoppe, F. M. (1987). The sampling theory of neutral alleles and an urn model in population genetics. *J. Math. Biol.* 25 123-159.

[8] Karlin, S. and McGregor, J. (1964). On some stochastic models in genetics. In *Stochastic Models in Medicine and Biology* (J. Gurland, ed.) 245-279. Univ. of Wisconsin Press, Madison.

[9] Kelly, F. P. (1977). Exact results for the Moran neutral allele model. *Adv. in Appl. Probab.* 9 197-201.

[10] Kingman, J. F. C. (1982a). On the genealogy of large populations. *J. Appl. Probab.* 19A 27-43.

[11] Kingman, J. F. C. (1982b). Exchangeability and the evolution of large populations. In *Exchangeability in Probability and Statistics* (G. Koch and F. Spizzichino, eds.) 91-112. North-Holland, Amsterdam.

[12] Kingman, J. F. C. (1982c). The coalescent. *Stochastic Process. Appl.* 13 235-248.

[13] Tavaré, S. (1989). The genealogy of the birth, death, and immigration process. In *Mathematical Evolutionary Theory* (M. W. Feldman, ed.) 41-56. Princeton U.P.

[14] Watterson, G. A. (1984). Lines of descent and the coalescent. *Theoret. Population Biol.* 26 77-92.

[15] Watterson, G. A. (1985). Estimating species divergence times using multilocus data. In *Population Genetics and Molecular Evolution* (T. Ohta and K. Aoki, eds.) 163-183. Japan Sci. Soc. Press, Tokyo/Springer-Verlag, Berlin.

[16] Watterson, G. A. (1987). Estimating the proportion of neutral mutants. *Genet. Res., Camb.* 501 155-163.

[17] Watterson, G. A. (1989). The neutral alleles model with bottlenecks. In *Mathematical Evolutionary Theory* (M. W. Feldman, ed.) 26-40. Princeton U. P.

Figure 1

A diagram showing part of a population history. Time runs down the page. At a birth, the parent is shown by a vertical line, the offspring by a slanting line. Mutant births are denoted by black circles. A sample of six individuals is denoted by black squares, and the sample-fores arising therefrom by thick lines.