Jour. Ind. Soc. Ag. Statistics
57 (Special Volume), 2004 : 84-93

On an Extension of the Theory of Coalescents to Populations with Two Sexes

Edward Pollak
Iowa State University, Ames, Iowa, USA

SUMMARY

It is assumed in this paper that there is a random mating population with two sexes, no selection or mutation, and an autosomal locus. In addition, there are in each generation the same large numbers N_m and N_f of males and females as in any other generation. It can then be shown that coalescence theory can be derived if the unit of measurement of time is taken to be $2N_e$ generations, where N_e is the effective population size when there are arbitrary offspring distributions.

Key words : n-coalescent, Dioecious random mating populations.

1. Introduction

The aim of coalescent theory is to infer features of the past history of a population from a sample taken at the present time. It is, of course, necessary to make assumptions about the mating system that produced individuals of each generation whether there was mutation, selection, recombination, or changing population size. The papers in which this theory was first developed in considerable generality are those of Kingman ([6], [7], [8]) and Tajima [11]. However, some aspects of it had been derived earlier by Felsenstein [2]. I shall first discuss a special case of his theory.

Let us assume that there are neutral alleles in a haploid population of constant size N, from which a random sample is drawn without replacement. Following Wright [12], I assume further that each generation arises from random sampling, with replacement, of parents. Then, if p_{kj} is the probability that k offsprings are derived from exactly j parental individuals one generation earlier then

$$p_{kk} = \left(1-\frac{1}{N}\right)\left(1-\frac{1}{N}\right)\ldots\left(1-\frac{k-1}{N}\right) \approx 1-\frac{k(k-1)}{2N} \tag{1}$$

$$p_{k,k-1} = \binom{k}{2}\frac{1}{N}\left[\left(1-\frac{1}{N}\right)\ldots\left(1-\frac{k-2}{N}\right)\right] \approx \frac{k(k-1)}{2N} \tag{2}$$

and

$$p_{k,k-r} = 0(N^{-2}) \quad \text{as } N \to \infty \tag{3}$$

It can be shown from equations (1) – (3) that if the unit of time is taken to be N generations, the time τ_k, it takes to reduce the number of ancestors from k to k – 1, has approximately an exponential distribution with the density function

$$f\tau_k(t) = \binom{k}{2} \exp\left[-\binom{k}{2} t\right], \; t > 0$$

Moreover, as the genealogy of a current sample of n individuals is traced back in time, successive reductions of numbers of ancestors from n to n – 1, n –1 to n – 2, ..., 2 to 1, are independently distributed.

Note, in addition, that the probability of any particular reduction from k to k – 1 ancestors in one generation, given such a reduction, is equal to $2/k(k-1)$. These are transition probabilities of what Kingman [6] refers to as the jump chain.

Kingman ([6], [7], [8]) studied genealogical relationships between pairs of individuals (x, y) (where x could be the same as y) of generation 0, with reference to past times s = –1, –2, ... Thus, for a specific s, x and y are in the same equivalence class if and only if they have a common ancestor in generation s. Let us denote the set of these equivalence equations by R_s and the number of equivalence classes by $|R_s|$. In other words, R_s contains all the ways that pairs of individuals are related with reference to generation s and $|R_s|$ denotes the number of ancestors at time s.

All distinct individuals at time 0 are defined to be not related. There are thus n equivalence classes, each with one relation, in generation 0. But, if one pair of individuals had a common parent, the number of equivalence classes is reduced from $|R_0| = n$ to $|R_1| = n - 1$, whereas the number of equivalence relations increases from n to n + 1. In general, going backward in time (as s decreases) is associated with an increase in the number of elements in R_s and a decrease in $|R_s|$. In the long run, as $s \to \infty$, all individuals in the present sample are related and there is only one common ancestor.

The theory just described has been extended in many directions. As far as I know, however, Möhle [9] is the only author who has considered a population of dioecious diploids. His paper has a complete theory for a generalization of the Wright model in which there are N couples in each generation.

My purpose in this paper is to sketch at least part of a coalescent theory for a random mating population with two sexes, an autosomal locus, no selection or mutation, and arbitrary offspring distributions. After an introduction of assumptions and notations in Section 2, rates of reduction in numbers of ancestors from 2 to 1 and from k to k – 1, $k \geq 3$, will be derived in Sections 3

and 4. In Section 5, I shall derive the distributions of times for such reductions. It will be assumed throughout this paper that the number of copies of a gene sampled in generation 0 is n.

Before proceeding, I note that if there are two sexes, separate copies of a gene in one individual must come from distinct parents. In a random mating population, these parents are a randomly chosen male-female pair, each contributing one copy of a gene to the offspring individual. A consequence will be that in generations -1, -2, ..., we need only to consider probabilities of sets of copies of a gene, each in a separate parent, if N_m and N_f are large.

2. Assumptions and Notations

Consider a random mating population that has large numbers, N_m and N_f, of males and females respectively in each generation. If, at time s, a random sample of distinct individuals is taken from the population and found to contain $k - i$ copies of a gene in males and i copies in females, the sample will be said to have configuration $(k - i, i)$. Now any pair of copies of a gene in the sample could have come either from random separate parents or from the same parent. The probabilities of the latter type of event are

$P_{u,vw}$ = P [two copies of a gene in random separate offspring of sexes v and w come from the same parent of sex u]

where u, v and w can each be replaced by either of the symbols m or f, denoting respectively, male or female.

It will be shown that these probabilities are small if N_m and N_f are large. Therefore, the configuration $(k - i, i)$ at time s is overwhelmingly likely to be derived from configurations of the types $(k - j, j)$ or $(k - 1 - j, j)$ at time $s - 1$. It follows that the matrix of probabilities of transition from states at time s to states at time $s - 1$ has approximately a structure with $k + 1 \times k + 1$ submatrices $\mathbf{Q}_k$ on the diagonal and, immediately to their left, submatrices $\mathbf{R}_k$ with $k + 1$ rows and k columns, for $k = 1, 2, \ldots, n$. These submatrices respectively have as elements probabilities of transitions that maintain the number of ancestors and probabilities of transitions that result in reductions of the number of ancestors by one.

3. The Rate of Reduction of the Number of Ancestors from 2 to 1

I list the possible configurations in the order (1, 0), (0, 1), (2, 0), (1, 1), (0, 2), and denote their probabilities in generation s by f((1, 0)s), f((0, 1)s), f((2, 0)s), f((1,1)s) and f((0, 2)s). The two copies of a gene in random separate offspring of sexes v and w in generation s come from (1, 0) or (0, 1) in generation $s - 1$ with probabilities $P_{m,vw}/8$ and $P_{f,vw}/8$, because there are 4 equally likely ways to choose the sexes of parents of two random separate copies and ½ is the conditional probability that two copies of a gene are derived

from a single parental copy if they both come from one particular parent. Thus, if $\mathbf{f}_s = [f((1, 0)s), f((0, 1)s), f((2, 0)s), f((1, 1)s), f((0, 2)s)]'$

$$\mathbf{f}_s = \begin{bmatrix} \mathbf{I} & \mathbf{0} \\ \mathbf{R}_2 & \mathbf{Q}_2 \end{bmatrix} \tag{4}$$

where

$$\mathbf{I} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$$

$$\mathbf{R}_2 = \frac{1}{8}\begin{bmatrix} P_{m,mm} & P_{f,mm} \\ P_{m,mf} & P_{f,mf} \\ P_{m,ff} & P_{f,ff} \end{bmatrix}$$

and

$$\mathbf{Q}_2 = \begin{bmatrix} \frac{1}{4}\left[1-\frac{1}{2}P_{m,mm}\right] & \frac{1}{2} & \frac{1}{4}\left[1-\frac{1}{2}P_{f,mm}\right] \\ \frac{1}{4}\left[1-\frac{1}{2}P_{m,mf}\right] & \frac{1}{2} & \frac{1}{4}\left[1-\frac{1}{2}P_{f,mf}\right] \\ \frac{1}{4}\left[1-\frac{1}{2}P_{m,ff}\right] & \frac{1}{2} & \frac{1}{4}\left[1-\frac{1}{2}P_{f,ff}\right] \end{bmatrix}$$

The matrix $\mathbf{Q}_2$ can be written as a sum of two matrices as follows

$$\mathbf{Q}_2 = \mathbf{Q}_{20} + \mathbf{Q}_{21}$$

where

$$\mathbf{Q}_{20} = \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix}\begin{bmatrix} \frac{1}{4} & \frac{1}{2} & \frac{1}{4} \end{bmatrix} = \mathbf{1}\mathbf{p}_2' \tag{5}$$

and

$$\mathbf{Q}_{21} = -\frac{1}{8}\begin{bmatrix} P_{m,mm} & 0 & P_{f,mm} \\ P_{m,mf} & 0 & P_{f,mf} \\ P_{m,ff} & 0 & P_{f,ff} \end{bmatrix} \tag{6}$$

This shows immediately that $\mathbf{Q}_{20}$ has one eigenvalue equal to 1 and corresponding left and right eigenvectors $\mathbf{p}_2'$ and $\mathbf{1}$, which satisfy the equation $\mathbf{p}_2'\mathbf{1} = 1$. Moreover, the other two eigenvectors are equal to 0. Since the probabilities $P_{u,vw}$ are all small, it follows from a standard result in perturbation theory as discussed, for example, by Franklin ([3], Section 6.12), that the dominant eigenvalue of $\mathbf{Q}_2$ is

$$\rho_2 = 1+\delta_2 \approx 1+\mathbf{p}_2'\mathbf{Q}_{21}\mathbf{1} = 1-\mathbf{p}_2'\mathbf{R}_2\mathbf{1}$$
$$= 1-\frac{1}{32}\left\{\left[P_{m,mm}+2P_{m,mf}+P_{m,ff}\right]+\left[P_{f,mm}+2P_{f,mf}+P_{f,ff}\right]\right\} \quad (7)$$

I now define G_{uv} to be the number of successful gametes contributed by a parent of sex u to adult offspring of sex v. If there is no selection

$$E(G_{uv}) = \frac{N_v}{N_u}$$

Since there is random mating, any pair of gametes in offspring of sexes v and w coming from parents of sex u is just as likely as any other pair. Thus

$$P_{u,vv} = \frac{N_u}{N_v(N_v-1)}E[G_{uv}(G_{uv}-1)] = \frac{1}{N_v-1}\left\{\frac{N_u}{N_v}\mathrm{Var}(G_{uv})+\frac{N_v}{N_u}-1\right\} \quad (8)$$

and

$$P_{u,mf} = \frac{N_u}{N_m N_f}E(G_{um},G_{uf}) = \frac{N_u}{N_m N_f}\mathrm{Cov}(G_{um},G_{uf})+\frac{1}{N_u} \quad (9)$$

Therefore, if the right sides of (8) and (9) are substituted for their corresponding left sides in (7), we obtain

$$\rho_2 \approx 1-\frac{1}{32}\left\{\left[\mathrm{Var}(G_{mm})+2\left(\frac{N_m}{N_f}\right)\mathrm{Cov}(G_{mm},G_{mf})+\left(\frac{N_m}{N_f}\right)^2\mathrm{Var}(G_{mf})+2\right]\right.$$
$$\left.+\frac{1}{N_f}\left[\mathrm{Var}(G_{ff})+2\left(\frac{N_f}{N_m}\right)\mathrm{Cov}(G_{ff},G_{fm})+\left(\frac{N_f}{N_m}\right)^2\mathrm{Var}(G_{fm})+2\right]\right\} \quad (10)$$

if N_m and N_f are large. Thus, $\rho_2 \approx 1-1/(2N_e)$, where N_e is the effective population size obtained by Caballero [1] and Nagylaki [10], who calculated probabilities of identity by descent in their derivations. It is also a special case of an expression for an age-structured population, derived by Hill ([4], [5]) in other ways.

4. The Rate of Reduction of the Number of Ancestors from k to k-1

If there is a sample of k copies of a gene, the distribution of the number of female parents from which the copies were derived is binomial with sample size k and probability of success ½. Thus, if there were an infinite population, the distribution of the number of female parents would be elements of the row vector

$$\mathbf{p}'_k = \frac{1}{2^k}\left[1\binom{k}{1}\binom{k}{2}\cdots\binom{k}{k-2}\binom{k}{k-1}1\right]$$

But if N_m and N_f are finite, we must also consider expressions

$$r\,[(k-i,\,i)|\,k-1]$$

= P [configuration (k – i, i) is derived from k – 1 distinct parent copies of a gene]

Any term of this type is a sum of elements in a row of $\mathbf{R}_k$ because it can be written as a sum over j of probabilities that one generation earlier there were configurations (k – 1 – j, j) followed by a transition to (k – i, i).

An equation for r [(k – i, i)| k – 1] will now be derived. First, note that there are $\binom{k-i}{2}$ ways to choose a pair of copies of a gene in k – i males, which are derived from a single copy one generation earlier. Given one such choice the probability is $P_{u,mm}/4$ that they were derived from a single gene in a parent of sex u. Thus, the total probability that two copies of a gene among males came from one parent copy in a parent of sex u is

$$\binom{k-1}{2}\frac{1}{4}P_{u,mm}$$

Likewise, the probability that two copies of a gene among i females are derived from a single copy in a parent of sex u is

$$\binom{i}{2}\frac{1}{4}P_{u,ff}$$

and the probability that two copies of a gene from a male-female pair came from a single gene in a parent of sex u is

$$(k-i)i\frac{1}{4}P_{u,mf}$$

Second, the parent with a gene that is copied twice is equally likely to be either a male or a female. Thus in summary, one may write

$$r\left[(k-i,i)\middle|k-1\right] = \frac{1}{8}\left\{\binom{k-i}{2}\left[P_{m,mm}+P_{f,mm}\right]+\binom{i}{2}\left[P_{m,ff}+P_{f,ff}\right]+(k-i)i\left[P_{m,mf}+P_{f,mf}\right]\right\} \tag{11}$$

Now note that

$$\mathbf{Q}_k = \mathbf{Q}_{k0} + \mathbf{Q}_{k1} = \mathbf{1}\mathbf{p}'_{k1} + \mathbf{Q}_{k1}$$

where $\mathbf{Q}_{k1}$ is a matrix whose elements are all at most of the order of magnitude $0(\max P_{u,vw})$. In addition, the sum of the elements in any row of $\mathbf{R}_k\,\mathbf{Q}_k$ is equal to 1 so that the sum of the elements of any row of $\mathbf{R}_k$ is minus the sum of elements in the same row of $\mathbf{Q}_{k1}$.

Just as when k = 2, the form of $\mathbf{Q}_{k0}$ immediately reveals that it has only one nonzero eigenvalue, which is equal to 1 and that corresponding left and right eigenvectors are $\mathbf{p}_k'$ and $\mathbf{1}$. Hence, by the same perturbation argument as used in the previous section, the dominant eigenvalue of $\mathbf{Q}_k$ is, by (11)

$$\begin{aligned}\rho_k &= 1+\delta_k \approx 1+\mathbf{p}_k'\mathbf{Q}_{k1}\mathbf{1} = 1-\mathbf{p}_k'\mathbf{R}_k\mathbf{1}\\ &= 1-\frac{1}{2^{k+3}}\sum_{i=0}^{k}\binom{k}{i}\left\{\binom{k-i}{2}\left[P_{m,mm}+P_{f,mm}\right]\right.\\ &\quad\left.+\binom{i}{2}\left[P_{m,ff}+P_{f,ff}\right]+(k-i)i\left[P_{m,mf}+P_{f,mf}\right]\right\}\end{aligned}$$

However

$$\frac{1}{2^k}\sum_{i=0}^{k}\binom{k}{i}\binom{k-i}{2}=\frac{k(k-1)}{8}\sum_{i=0}^{k}\frac{(k-2)!}{i!(k-1-2)!}\frac{1}{2^{k-2}}=\frac{k(k-1)}{8}$$

$$\frac{1}{2^k}\sum_{i=0}^{k}\binom{k}{i}\binom{i}{2}=\frac{k(k-1)}{8}\sum_{i=0}^{k}\frac{(k-2)!}{(k-i)!(i-2)!}\frac{1}{2^{k-2}}=\frac{k(k-1)}{8}$$

and

$$\frac{1}{2^k}\sum_{i=0}^{k}\binom{k}{i}i(k-i)=\frac{k(k-1)}{4}\sum_{i=0}^{k}\frac{(k-2)!}{(i-1)!(k-i-1)!}\frac{1}{2^{k-2}}=\frac{k(k-1)}{4}$$

Hence

$$\begin{aligned}\rho_k &\approx 1-\frac{1}{32}\binom{k}{2}\left\{\left[P_{m,mm}+2P_{m,mf}+P_{m,ff}\right]+\left[P_{f,mm}+2P_{f,mf}+P_{f,ff}\right]\right\}\\ &= 1-\binom{k}{2}\frac{1}{2N_e}\end{aligned} \tag{12}$$

5. Distribution of Times to Coalescence

If there is a reduction of the number of ancestors from k to k – 1 between generations s and s – 1, there will be said to be a coalescence event. It was shown in Section 4 that the rate of this reduction per generation is approximately $1-\binom{k}{2}\Big/(2N_e)$. The matrix $\mathbf{Q}_k^r$ has as its elements probabilities that there has not been such a reduction in r generations. It is approximately equal to

$$\mathbf{Q}_{k0}^{r} \approx \left[1-\binom{k}{2}\frac{1}{2N_e}\right]^{r}\mathbf{1p}_k'$$

which has elements that are independent of any initial configuration (k – i, i). So if T_k denotes the number of generations until the number of ancestral copies is reduced from k to k – 1

$$P[T_k \le r] \approx 1-\left[1-\binom{k}{2}\frac{1}{2N_e}\right]^{r}$$

If, therefore, we follow Kingman ([6], [7], [8]) and change the scale of measurement so that time is measured in units of $2N_e$ generations

$$P\left[\tau_k = \frac{T_k}{2N_e} \le t\right] \approx 1-\exp\left[-\binom{k}{2}t\right] = F_{\tau_k}(t)$$

Hence, T_k is approximately exponentially distributed with the density function

$$f_{\tau_k}(t) = \binom{k}{2}\exp\left[-\binom{k}{2}t\right], t>0 \qquad (13)$$

It follows from (13) that the mean and variance of this distribution are respectively equal to

$$E(\tau_k) = 1\Big/\binom{k}{2}$$

and

$$Var(\tau_k) = 1\Big/\binom{k}{2}^2$$

In terms of time measured in generations we obtain

$$E(T_k) \approx 2N_e\Big/\binom{k}{2}$$

and

$$Var(T_k) \approx 4N_e^{\,2}\Big/\binom{k}{2}^2$$

which is consistent with results obtained by Tajima [11], who studied an idealized monoecious population for which $N_e = N$, the actual population size.

Finally, another conclusion follows from the fact that we have a Markov chain. Thus, states at all times farther in the past than when there is a reduction from k to k – 1 ancestors are independent of what occurred up to the reduction. Consequently, the set of coalescence times $\tau_n, \tau_{n-1}, \ldots, \tau_2$ are independent exponentially distributed random variables.

6. Discussion

Each of the $\binom{k}{2}$ ways, a reduction from k to k – 1 ancestors can occur, has the same probability 2/k(k – 1). These are transition probabilities of the jump chain. Also, the number of ancestors at time t, with time measured in units of $2N_e$ generations, is a continuous time Markov chain with transition rates $\binom{n}{2}, \binom{n-1}{2}, \ldots, \binom{2}{2}$. This pure death process describes the timing of downward jumps in numbers of ancestors. More generally, however, if the sample at time 0 consists of n copies of a gene taken from separate individuals, we have a Markov chain whose state space is the set of equivalence relations on {1, 2, ..., n}, which is called the n-coalescent.

Given a downward jump in generation s, the past and future of the coalescent chain contain independently distributed times at which the number of ancestors stay constant and are unaffected by which particular jumps are made. Therefore, the jump chain is independent of the death process.

One apparent loose end in the theory described in this paper is the possible need to assume the sample observed at time 0 consists of n copies of a gene chosen from n separate individuals. Suppose, however, that there are r pairs of copies of a gene in r individuals in generation 0. Then, in generation -1, the parent copies are in r male-female pairs. Therefore, if N_m and N_f are large, the overwhelmingly most likely events in generation -1 are to have configurations (n – i, i) or, if one pair of copies in separate individuals comes from a common parent copy, configurations (n – 1 – i, i). Thus, if time is measured in units of $2N_e$ generations, the results obtained in this paper are unaffected by whether or not r individuals each contributed two separate copies of a gene to the sample at time 0.

REFERENCES

[1] Caballero, A. (1995). On the effective size of populations with separate sexes with particular reference to sex-linked genes. *Genetics*, **139**, 1007-1111.

[2] Felsenstein, J. (1971). The rate of loss of multiple alleles in finite haploid populations. *Theor. Popul. Biol.*, **2**, 391-403.

[3] Franklin, J.N. (1968). *Matrix Theory*. Prentice-Hall, Engelwood Cliffs, NJ.

[4] Hill, W.G. (1972). Effective size of populations with overlapping generations. *Theor. Popul. Biol.*, **3**, 278-289.

[5] Hill, W.G. (1979). A note on effective population size with overlapping generations. *Genetics*, **92**, 317-322.

[6] Kingman, J.F.C. (1982). The coalescent. *Stochastic Processes and Their Applications*, **13**, 235-248.

[7] Kingman, J.F.C. (1982). On the genealogy of large populations. In: *Journal of Applied Probability*, 19A, Essays in Statistical Science, Papers in Honor of P.A.P. Moran (J. Gani, E.J. Hannan eds.), 28-43.

[8] Kingman, J.F.C. (1982). Exchangeability and the evolution of large populations. In: *Exchangeability in Probability and Statistics* (G. Koch, F. Spizzichino eds.), 97-112. North Holland Publishing Company.

[9] Möhle, M. (1998). Coalescent results for two-sex population models. *Adv. Appl. Prob.*, **30**, 513-520.

[10] Nagylaki, T. (1995). The inbreeding effective population number in dioecious populations. *Genetics*, **139**, 473-485.

[11] Tajima, F. (1983). Evolutionary relationship of DNA sequences in finite populations. *Genetics*, **105**, 437-460.

[12] Wright, S. (1931). Evolution in Mendelian populations. *Genetics*, **16**, 97-159.