Genet. Res., Camb. (1991), **58**, *pp.* 167–175 *With* 1 *text-figure Printed in Great Britain*

Inbreeding coefficients and coalescence times

MONTGOMERY SLATKIN
Department of Integrative Biology, University of California, Berkeley CA 94720

(*Received 15 August 1990 and in revised form 1 May 1991*)

Summary

This paper describes the relationship between probabilities of identity by descent and the distribution of coalescence times. By using the relationship between coalescence times and identity probabilities, it is possible to extend existing results for inbreeding coefficients in regular systems of mating to find the distribution of coalescence times and the mean coalescence times. It is also possible to express Sewall Wright's F_{ST} as the ratio of average coalescence times of different pairs of genes. That simplifies the analysis of models of subdivided populations because the average coalescence time can be found by computing separately the time it takes for two genes to enter a single subpopulation and time it takes for two genes in the same subpopulation to coalesce. The first time depends only on the migration matrix and the second time depends only on the total number of individuals in the population. This approach is used to find F_{ST} in the finite island model and in one- and two-dimensional stepping-stone models. It is also used to find the rate of approach of F_{ST} to its equilibrium value. These results are discussed in terms of different measures of genetic distance. It is proposed that, for the purposes of describing the amount of gene flow among local populations, the effective migration rate between pairs of local populations, $\hat{M}$, which is the migration rate that would be estimated for those two populations if they were actually in an island model, provides a simple and useful measure of genetic similarity that can be defined for either allozyme or DNA sequence data.

1. Introduction

Population genetic models of neutral genes in finite populations have traditionally been developed in terms of inbreeding coefficients or probabilities of identities by descent. These models are appropriate for analyzing gene frequency data, from which inbreeding coefficients or identity probabilities can be estimated. More recently, a different class of model, called variously 'retrospective', 'genealogical' or 'coalescent' models has been introduced. These models differ from traditional models in focusing on the times at which two or more genes have a common ancestor in the past. Tavaré (1984), Ewens (1990) and Hudson (1990) provide reviews of coalescent models and their applications. These two classes of models are, of course, equivalent because they both describe the mathematical consequences of inheritance, mutation and genetic drift but they call on somewhat different mathematical tools and relate to different kinds of data. In this paper I show the relationship between these two classes of models, particularly in their application to subdivided populations. I will also show that by using coalescent models it is possible to derive some general results for subdivided populations because the effects of different processes can be analysed separately.

2. Preliminaries

(i) *Identities by descent and coalescence times*

Throughout, I will be concerned with pairs of neutral genes at a single locus. In this context, a gene is a non-recombining segment of DNA. For our purposes, inbreeding coefficients, as introduced by Wright (1922), and probabilities of identity by descent, as defined by Malécot (1948) are equivalent, although Wright (1969, p. 178) emphasized that they are not always so because inbreeding coefficients can be negative while probabilities of identity by descent cannot. For simplicity, I will use the term probability of identity and denote probabilities of identity at time t by $f(t)$, which will have different subscripts to indicate which pair of genes is being considered.

The results derived here will depend on the

relationship between coalescence times and genetic identity. Define $g(t)$ to be the probability that two genes do not have a common ancestor by generation t in the past or equivalently the probability that two genes have not coalesced by generation t in the past. Let $P(t)$ be the probability that the coalescence occurs in generation $t (t \geqslant 1)$. Clearly

$$P(t) = g(t-1) - g(t). \tag{1}$$

Note that $\sum_{t=1}^{\infty} P(t) = g(0) = 1$ because, in a finite population, $g(t)$ must go to 0 for large t. It is also useful to note that the mean coalescence time is

$$\bar{t} = \sum_{t=1}^{\infty} tP(t) = \sum_{t=0}^{\infty} g(t). \tag{2}$$

The reason for introducing $g(t)$ is that the recursion equations describing $g(t+1)$ as a function of $g(t)$ are exactly the same as the recursion equations describing the probability of non-identity, $1-f(t+1)$, as a function of $1-f(t)$ in the absence of mutation, which can be seen by noting that the probability of non-identity decreases only if a coalescence event occurs. The equivalence of these two quantities means that existing theories describing changes in probabilities of identities by descent can be used with only minor changes to predict the distribution of coalescence times and average coalescence times.

(ii) *Regular systems of mating*

We can see the utility of relating the probability of non-identity to $g(t)$ by considering regular systems of mating. Here, I will discuss only systems in which there is the same number of individuals in each generation. Consider, for example, a system of regular full-sib mating. In this case, two probabilities of identity are needed, $f_I(t)$, the probability that the two homologous genes from the same individual are identical, and $f_{IJ}(t)$, the probability that genes from different individuals are identical. For continued full-sib mating, two linear recursion equations are needed to determine $g_I(t)$ and $g_{IJ}(t)$ and these can be obtained from the standard recursion equations for $f_I(t)$ and $f_{IJ}(t)$ (Crow and Kimura 1970, p. 87). The resulting equations can be written in matrix form as

$$\mathbf{g}(t+1) = A\mathbf{g}(t), \tag{3}$$

where

$$\mathbf{g}(t) = \begin{pmatrix} g_I(t) \\ g_{IJ}(t) \end{pmatrix},$$

$$A = \begin{pmatrix} 0 & 1 \\ \frac{1}{4} & \frac{1}{2} \end{pmatrix}. \tag{4}$$

Equation (1) has the solution $\mathbf{g}(t) = A^t\mathbf{g}(0)$ with $g_I(0) = g_{IJ}(0) = 1$. The distribution of coalescence times, $\mathbf{P}(t) = \mathbf{g}(t-1) - \mathbf{g}(t)$ can be found directly. The mean coalescence times can also be found,

$$\bar{t} = \begin{pmatrix} \bar{t}_I \\ \bar{t}_{IJ} \end{pmatrix} = \sum_{t=0}^{\infty} A^t\mathbf{g}(0) = (I-A)^{-1}\mathbf{g}(0). \tag{5}$$

Obviously, this approach can be applied to any regular system of mating in which the same number of individuals are present each generation. The only changes in the above calculations are that more values of $g(t)$ are needed and the matrix A has to be redefined. Then, (3) and (5) provide the necessary results.

(iii) F_{ST} *and mean coalescence times*

Wright (1951) partitioned the inbreeding coefficient of an individual relative to a collection of local populations, F_T, into a part attributable to non-random mating within a local population, F_{IS}, and a part attributable to differences among populations, F_{ST}. The quantity F_{ST} can be expressed in terms of probabilities of identity

$$F_{ST} = \frac{f_0 - \bar{f}}{1-\bar{f}}, \tag{6}$$

where f_0 is the probability of identity of two genes sampled from the same subpopulation and $\bar{f}$ is the probability of identity of two genes sampled at random from the collection of subpopulations being considered (Nei, 1973). It is possible to compute probabilities of identity in various models of population structure and hence predict the values of F_{ST} under different conditions (Slatkin & Barton, 1989). The problem with these calculations is that in general, both f_0 and $\bar{f}$ depend not only on effective population sizes and the migration matrix but also on the intensity of mutation or other force that is maintaining genetic variation. It would be desirable to predict the value of F_{ST} in a way that did not confound the purely demographic processes of genetic drift and migration with purely genetic processes such as mutation. We can do this by expressing F_{ST} in terms of coalescence times.

If we assume that mutation is a Poisson process, then the probability of identity of two genes is the probability that no mutation occurred before those genes had a common ancestor (Hudson, 1990). We can compute the value of f for any pair of genes in terms of coalescence times by noting that the probability that a mutation has not occurred by generation t is $(1-\mu)^t$ where μ is the mutation rate. Therefore, the probability that two genes are identical is

$$f = \sum_{t=1}^{\infty} (1-\mu)^{2t} P(t), \tag{7}$$

where $P(t)$ is the probability that the two genes coalesced in generation t in the past. Clearly, f considered as a function of μ goes to 1 as μ goes to 0. However, F_{ST} defined by equation (6) has a non-

https://doi.org/10.1017/S0016672300029827 Published online by Cambridge University Press

trivial limit as μ goes to 0, which is found by applying l'Hôpital's rule:

$$F_{ST} = \frac{\bar{t} - \bar{t}_0}{\bar{t}}. \tag{8}$$

where $\bar{t}_0$ is the average coalescence time of two genes drawn from the same subpopulation and $\bar{t}$ is the average coalescence time of two genes drawn from the collection of subpopulations. Equation (8) can also be obtained by noting that for small μ, $f \approx 1 - \mu\bar{t}$.

Equation (8) allows us to express F_{ST} in a way that does not depend on the mutation rate if mutation is a weak force. Furthermore, if the DNA sequence of genes are known, then those sequences can be used to estimate the $\bar{t}$s directly and hence F_{ST}, which is in effect what Nei (1982) suggests. In what follows, I will find values of F_{ST} in different models of population structure by computing average coalescence times and then using equation (8).

The value of F_{ST} is often used to estimate average levels of gene flow. Wright (1951) showed that at equilibrium in an island model of population structure

$$Nm \approx \frac{1}{4}\left(\frac{1}{F_{ST}} - 1\right), \tag{9}$$

where m is the migration rate and N is the effective population size. If we denote the estimate of Nm obtained from equation (9) by $\hat{M}$, then there is a simple relationship between $\hat{M}$ and the average coalescence times.

$$\hat{M} = \frac{\bar{t}_0}{4(\bar{t} - \bar{t}_0)}. \tag{10}$$

(iv) *Average coalescence times in a subdivided population*

In equations (8) and (10), the value of $\bar{t}_0$, the average coalescence time of two genes in the same subpopulation, enters. That is especially convenient because the value of $\bar{t}_0$ is, under a very general model of migration, independent of the migration pattern and depends only on the total number of individuals in the population being considered. Strobeck (1987) considered a model of a population divided into d subpopulations, with the ith subpopulation containing N_i individuals. He assumed that migration among the subpopulations was isotropic and conservative, and considered the infinite sites model of mutation, in which each site can change only once. He showed that the expected number of sites that differ between two randomly chosen genes from the same population is $4N_T\mu$ where $N_T = \Sigma N_i$ and μ is the mutation rate. Because mutation in this model is a Poisson process the expected number of differences is $2\bar{t}_0\mu$ and hence $\bar{t}_0 = 2N_T$, independently of the elements of the migration matrix. Hey (1991) obtained the same result using the general theory of Markov chains. Takahata (1988) analyses models of more than two genes in population containing two demes.

This result greatly simplifies the calculation of F_{ST} from (8). The value of $\bar{t}$ depends on the model of population structure being considered and on which populations are sampled. The value of $\bar{t}$ will be the sum of two terms, with one term representing genes that are in the same subpopulation and one term representing genes that are in different subpopulations. We already know the contribution of genes in the same subpopulation because that is $\bar{t}_0 = 2N_T$. Genes that are in different subpopulations cannot coalesce until they are present in the same subpopulation, after which the mean coalescence time is again $\bar{t}_0$. Thus, all that is needed to compute F_{ST} and $\hat{M}$ in any model of a subdivided population is the average time it takes for two genes present in different subpopulations to be first present in the same population. That problem is a standard problem in the theory of Markov processes and one that has already been solved in numerous special cases. The fact that we are concerned with the process going backwards rather than forward in time makes no difference in the calculation as long as we use the appropriate backwards migration matrix.

3. Populations at equilibrium

(i) *Finite island model*

In the finite island model, there are d demes each containing N individuals. Hence, $N_T = dN$ and $\bar{t}_0 = 2dN$. If we sample two genes at random from the set of d demes, there is a probability $(1/d)$ that both will be drawn from the same deme and $(d-1)/d$ that they will be drawn from different demes. Therefore,

$$\bar{t} = \frac{1}{d}\bar{t}_0 + \frac{d-1}{d}\bar{t}_1, \tag{11}$$

where $\bar{t}_1$ is the average time until two genes drawn from different demes coalesce.

Consider two genes that are in different demes at time t in the past. They will be present in the same deme at $t+1$ only if either one of them has emigrated from the deme the other occupies or if they both have emigrated from the same deme. Let m be the probability that each gene is an emigrant from some other deme. The probability that one gene did not emigrate in the previous generation is $1-m$ and the probability that the other emigrated from that same deme is $m/(d-1)$. The probability that both emigrated from the same deme is $m^2/(d-1)$. Hence, the net probability that the two genes were in the same deme in the preceding generation is

$$\frac{2m(1-m)}{d-1} + \frac{m^2}{d-1} \approx \frac{2m}{d-1},$$

if m is small. Therefore the average time until two genes are present on the same deme is approximately

https://doi.org/10.1017/S0016672300029827 Published online by Cambridge University Press

$(d-1)/(2m)$. Once they are present in the same deme, their average coalescence time is $\bar{t}_0 = 2Nd$. Therefore

$$\bar{t}_1 = 2Nd + \frac{d-1}{2m}, \tag{12}$$

and

$$\bar{t} = 2Nd + \frac{(d-1)^2}{2dm}, \tag{13}$$

so (8) implies

$$F_{ST} = \frac{1}{1+4Nmd^2/(d-1)^2}, \tag{14}$$

a result that was obtained by Takahata (1983) and Crow & Aoki (1984) under the assumption of small mutation rates.

(ii) *Stepping-stone models*

The value of F_{ST} in a stepping-stone model of population structure is in general a function of the locations of the demes that are sampled and the mutation rate (Slatkin & Barton, 1989). We can use the coalescent approach to find values of F_{ST} for samples taken from pairs of demes a specified distance apart. Consider first a circle of d demes with migration rate $m/2$ between adjacent demes and assume that d is even. A circular array simplifies the analysis because we can assume that only the separation of demes sampled is important. The same approach can be used for a linear array. Assume that a value of F_{ST} is computed for samples from only two demes separated by i steps ($i = 0, 1, \ldots, d/2$). According to equation (8),

$$F_{ST}(i) = \frac{\bar{t}_i - \bar{t}_0}{\bar{t}_i + \bar{t}_0}, \tag{15}$$

where $\bar{t}_i$ is the average coalescence time for two genes sampled from demes i steps apart, because $\bar{t} = (\bar{t}_i + \bar{t}_0)/2$ in this case.

As in the island model, $\bar{t}_i$ is the sum of two parts, the average time for two genes to occupy the same deme and the average time to coalesce given that they occupy the same deme. The second time is $\bar{t}_0 = 2Nd$. The first time can be found from the standard theory of random walks (Feller, 1957, ch. 14). The average time until two genes i steps apart initially are first found in the same deme is $(d-i)i/2m$ so

$$\bar{t}_i = 2Nd + \frac{(d-i)\,i}{2m} \tag{16}$$

and

$$F_{ST}(i) = \frac{1}{1+8Nmd/[(d-i)\,i]}. \tag{17}$$

If d is large and $i \ll d$, then

$$F_{ST}(i) = \frac{1}{1+8Nm/i} \tag{18}$$

and (10) implies that $\hat{M}(i) = 2Nm/i$ when $i \ll d$.

In a two-dimensional stepping-stone model, the results are similar but more difficult to obtain. Consider a two-dimensional 'torus' containing d demes in each direction, each of size N. In this case $N_T = Nd^2$. Assume that in each generation, a gene remains in its deme with probability $1-m$ and emigrates to one of the four adjacent demes with probability $m/4$. Let $\bar{t}_{ij}$ be the mean coalescence time of two genes sampled from i and j steps apart in the two directions ($i, j = 0, \ldots, d/2$). We know that $\bar{t}_{00} = \bar{t}_0 = 2Nd^2$. To compute $\bar{t}_{ij}$, we need to solve a standard problem in the theory of random walks on a torus. The results are derived in the Appendix. For small values of m, the results are especially convenient:

$$F_{ST}(i,j) = \frac{1}{1+32Nm/S(i,j)}, \tag{19a}$$

and hence

$$\hat{M} = \frac{8Nm}{S(i,j)}, \tag{19b}$$

where $S(i,j)$ is a function whose value depends on only i, j, and d [see equation (A 10) in the Appendix]. Although S does not appear to be expressible in a simple form, it is easy to evaluate numerically. The Appendix shows that if d is large, $S(i,0)$ is approximately independent of d.

The results are very similar in character to those obtained by Slatkin & Maddison (1990). In that paper, we used a cladistic method to estimate Nm from a cladogram of genes sampled from a subdivided population. We denoted the estimate of Nm obtained by assuming an island model of population structure by $\hat{M}$. The definition of $\hat{M}$ that follows from equation (9) is exactly the same: it is the estimate of Nm obtained by assuming an island model. Using extensive simulations, Slatkin & Maddison (1990) found that $\hat{M}$ was approximately proportional to the inverse of distance in a one-dimensional stepping-stone model and was approximately proportional to the inverse of the square root of distance in a two-dimensional stepping-stone model. Equation (18) implies that $\hat{M}$ is defined here is also inversely proportional to distance in a one-dimensional model.

The results for the two-dimensional model are also similar to those of Slatkin & Maddison (1990). Fig. 1 shows some numerical results obtained by summing the series defining $S(i,0)$ in (19) [and also showing the integral approximation, equation (A 12)]. The left hand graph shows the numerical results plotted on a long-linear scale, on which a straight line indicates logarithmic dependence of $S(i,0)$ on i. The right hand

https://doi.org/10.1017/S0016672300029827 Published online by Cambridge University Press

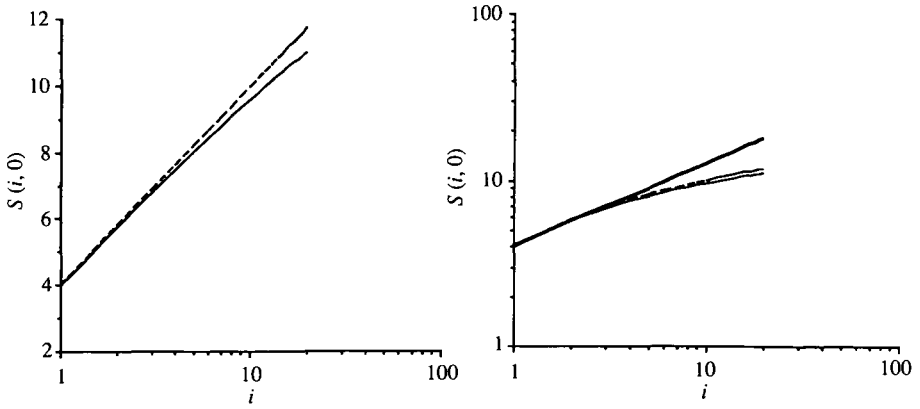

Fig. 1. Values of $S(i,0) = \Sigma'/d^2$, where Σ' is sum in equation (A 8). In both graphs, the thin solid lines show the exact values obtained from the sum, and the dashed lines show the approximate results obtained by numerically evaluating the integral in equation (A 12). In the right-hand graph, the heavier solid line is a graph of $4\sqrt{i}$.

graph shows the same results plotted on a log-log scale and compares the numerical results with a plot of $4\sqrt{i}$. Clearly, these results are best fit by a logarithmic function of i, but for small i, they are also close to the square root dependence found by Slatkin & Maddison (1990).

We can conclude that in a model of isolation by distance, the effective migration rate obtained by assuming an island model, $\hat{M}$, is an appropriate measure of genetic differentiation if the goal is to determine the extent and pattern of gene flow. How $\hat{M}$ is defined depends on the kind of data available and the way it is analysed.

4. Rate of approach to equilibrium

The results so far has assumed that the population of interest is at an equilibrium under migration and genetic drift. A disturbing feature of these results is that values of F_{ST} computed from equation (8) depend on the ratio of average coalescence times that may in fact be very long, possibly much longer than a population could remain at equilibrium under realistic conditions. For that reason, it is of interest to understand what determines the rate of approach of F_{ST} to its equilibrium value.

Assume that we are concerned with a subdivided population that has been present for τ generations. At time τ in the past, individuals forming the population were drawn at random from a single population. The question of interest is the extent of genetic identity that accumulates after the population is subdivided, so it is reasonable to assume that all genes are not identical by descent in the population at $t = \tau$. Define $\tilde{P}(t,\tau)$ to be the probability that the two genes coalesce in generation t given that they coalesce before τ and, as before, $g(t)$ is the probability that two genes have not coalesced before generation t. Clearly

$$\tilde{P}(t,\tau) = P(t)/\sum_{t=1}^{\tau} P(t) = P(t)/[1-g(\tau)]. \tag{20}$$

The probability that two genes are identical is

$$f = \sum_{t=1}^{\tau} (1-\mu)^{2t}\tilde{P}(t,\tau). \tag{21}$$

If we now consider F_{ST} in a population not at an equilibrium and use (21) to define the appropriate f's, we find that in the limit as μ goes to 0,

$$\tilde{F}_{ST} = \frac{\tilde{t}-\tilde{t}_0}{\tilde{t}}, \tag{22}$$

where $\tilde{t}_0$ is the average coalescence time of two genes in the same subpopulation given that they coalesce before τ and $\tilde{t}$ is the average coalescence time of two genes chosen at random from the collection of subpopulations given that they coalesce before τ. $\tilde{F}_{ST}$ defined by (22) depends on τ. As τ increases, $\tilde{F}_{ST}$ approaches F_{ST}.

Equation (22) allows us to compute F_{ST} for a population that has not had time to reach an equilibrium under the balance between migration and genetic drift. To illustrate how $\tilde{t}_0$ and $\tilde{t}$ are found, consider the finite island model. Let $g_0(t)$ be the probability that two genes initially in the same deme have not coalesced by generation t and let $g_1(t)$ be the probability that two genes initially in different demes have not coalesced by generation t. Going backwards in time, the transition probabilities for $g_0(t)$ and $g_1(t)$ can be written in matrix form:

$$\mathbf{g}(t+1) = A\mathbf{g}(t), \tag{23}$$

where

$$\mathbf{g}(t) = \begin{pmatrix} g_0(t) \\ g_1(t) \end{pmatrix} \tag{24a}$$

and

$$A = \begin{pmatrix} 1-2m-1/(2N) & 2m \\ 2m/(d-1) & 1-2m/(d-1) \end{pmatrix} \tag{24b}$$

where, for simplicity, we assume that m and $1/(2N)$

https://doi.org/10.1017/S0016672300029827 Published online by Cambridge University Press

are both small, in which case only one transition can occur in any generation.

Equation (23) is the same equation that arose in considering regular systems of mating and has the same solution: $\mathbf{g}(t) = A^t\mathbf{g}(0)$, with the relevant initial condition, $g_0(0) = g_1(0) = 1$. From equation (20) it is clear that $\tilde{P}(t)$ will be nearly $P(t)$ if $g(\tau)$ is small. Therefore, the rate of approach of $\bar{F}_{ST}$ to its equilibrium value is determined by the rate of approach of $g_0(t)$ and $g_1(t)$ to zero, which is in turn determined by the largest eigenvalue of A,

$$\lambda = 1-\frac{1}{4N}-\frac{m}{d-1}+\frac{1}{4N}\sqrt{\left(1+\frac{16N^2m^2d^2}{(d-1)^2}+\frac{8Nm(d-2)}{d-1}\right)}. \quad (25)$$

Roughly speaking, the rate of approach of F_{ST} to its equilibrium value is governed by the smaller of $1/N$ and m/d. If $\tau \gg 1/\lambda$ then F_{ST} will be approximately at its equilibrium value which is given by equation (14).

This method can be used with any other model of a subdivided population. The matrix A needs to be redefined to describe the particular migration pattern assumed. That allows us to use existing results concerning the rate of loss of genetic variability in a subdivided population to predict the rate at which F_{ST} approaches equilibrium. For example, in a circle of demes, Maruyama (1970*a*) showed that the dominant eigenvalue of A is approximately $1-(\bar{m}\pi^2)/(2d^2)$ if $Nm < d/10$ and is approximately $1-1/(4Nd)$ if $Nm > d/10$. For a two-dimensional array of demes on a torus, Maruyama (1971) stated that $\lambda \approx 1-1/(2Nd)$ if $Nm > 1$ and $\lambda \approx 1-2m/d$ if $Nm < 1$. Maruyama (1971) discussed the qualitative difference between the results for one and two-dimensional stepping-stone models.

5. Discussion

(i) *Estimators of* F_{ST} *from DNA sequence data*

DNA sequence data can be used to estimate F_{ST} because differences in sequence can be used to estimate divergence times of pairs of genes. There are several statistics proposed for the analysis of sequence and restriction site data that are related to F_{ST}. Nei (1982) proposed the statistic γ_{ST} to characterize nucleotide diversity within and between subpopulations. Takahata & Palumbi (1985) proposed computing Nei's (1973) G_{ST} by treating each polymorphic restriction site as a single locus. Their procedure computing G_{ST} is equivalent to that for computing Nei's (1982) γ_{ST}. Lynch & Crease (1990) propose a closely related statistic, N_{ST}.

All of these statistics depend on nucleotide diversities within subpopulations, $\hat{v}_w$, and between subpopulations, $\hat{v}_b$ [following Lynch & Crease's (1990) notation]. They all have the form $\hat{v}_b/(\hat{v}_w+\hat{v}_b)$. The difference between γ_{ST} and N_{ST} is that, for γ_{ST}, the value of $\hat{v}_b$ includes that possibility that two genes may be in the same subpopulation [cf. equation (11)] while for N_{ST} it does not. The values of $\hat{v}_w$ and $\hat{v}_b$ provide estimators of average coalescence times if mutation is assumed to be a Poisson process. In fact, $\hat{v}_w = 2\mu\bar{t}_0$ and $\hat{v}_b = 2\mu(\bar{t}-\bar{t}_0)$, if we follow Nei's (1982) definition of $\hat{v}_b$. Hence γ_{ST} estimates F_{ST} as defined by equation (8).

(ii) *Estimating average levels of gene flow from DNA sequence data*

Values of F_{ST} estimated from allozyme data are commonly used to estimate the average level of gene flow (Nm) by using equation (9). In doing so, some method is chosen for combining information from different loci to produce a single estimate of F_{ST} and hence Nm (Slatkin & Barton, 1989). It is clear that the same approach can be used with DNA sequence data. Nei's (1982) γ_{ST} or Lynch & Crease's (1990) N_{ST} can be computed for a particular data set and then used to estimate Nm. The problem with this approach is that the resulting estimates of Nm will not be very accurate because currently available studies provide sequences of only one part of the genome, usually the mitochondrial genome which for the purposes here is a single gene. Furthermore, statistics such as γ_{ST} and N_{ST} do not make full use of the information in the data. DNA sequences provide more than just information about coalescence times between different pairs of genes. The sequences also provide information about the gene tree. Maddison and I (Slatkin & Maddison, 1989) have shown that a method that uses some of this phylogenetic information provides estimates of Nm that are comparable in accuracy to estimates made using F_{ST} based on 10 allozyme loci. Unpublished simulations by Hudson, Slatkin and Maddison show that Slatkin & Maddison's (1989) cladistic method yields much more accurate estimates of Nm than does F_{ST} if there is no recombination.

(iii) *Inferring patterns of gene flow*

There have been a variety of statistics proposed to estimate genetic distances or similarities. Nei (1987) reviews several statistics that have been developed for both gene frequency and DNA sequence data. The approach taken here can show the relationship between different statistics and average coalescence times and suggest an appropriate measure of genetic similarity if the goal is to understand the pattern of gene flow.

One measure of genetic distance is the value of F_{ST} for a pair of populations. We used that measure to describe the results for stepping-stone models. For any pair of populations,

$$F_{ST} = \frac{\bar{t}_1-\bar{t}_0}{\bar{t}_1+\bar{t}_0}, \quad (26)$$

https://doi.org/10.1017/S0016672300029827 Published online by Cambridge University Press

where $\bar{t}_0$ the average coalescence time for two genes drawn at random from the same population and $\bar{t}_1$ is the average coalescence time for two genes, one drawn from each population. Thus, F_{ST} measures the increase in the mean coalescence time attributable to genes being in different population relative to the average coalescence time.

We can contrast F_{ST} as a measure of genetic distance with Nei's (1972) distance, D, which is related to identities by descent by $D = -\ln(f_1/f_0)$, where f_0 is the probability of identity of two genes drawn from the same population and f_1 is the probability of identity of two genes drawn from different populations. In the limit of small mutation rates, equation (7) implies that $f_0 \approx 1-\mu\bar{t}_0$ and $f_1 \approx 1-\mu\bar{t}_1$ and hence

$$D \approx \mu(\bar{t}_1 - \bar{t}_0). \tag{27}$$

If D is measured between populations that have been separated for a long time, then $\bar{t}_1$ is approximately the time of separation of those populations, in which case, $\bar{t}_1 \gg \bar{t}_0$ and $D \approx \mu\bar{t}_1$. That result is already known (Nei, 1972), namely that D increases in proportion to the time of separation of two populations, if mutation rates are small. I have rederived it here to show the relationship between F_{ST} and D as measures of genetic distance.

The value of F_{ST} depends on the difference in coalescence times scaled by the mean coalescence time while D depends on the difference in coalescence times scaled by the mutation rate. Hence, for small mutation rates, F_{ST} depends only on those factors that make coalescence times different, factors such as migration patterns, population densities and breeding systems. In contrast, D necessarily depends both on those factors and on mutation rates. As Nei (1987) discusses, other genetic distances tend to be similar in character to either F_{ST} or D so other distances can be easily related to average coalescence times.

Which measure of genetic distance is appropriate depends on what use is made of the results. If data are sampled from different species and the goal is to estimate divergence times of those species, then Nei's or a genetic distance with similar properties is appropriate because such distances are, under biologically reasonable assumptions, proportional to times of separation. If, on the other hand, data are sampled from a single species and the concern is with the extent and pattern of gene flow, then the measure of genetic distance should be chosen to reflect the amount of gene flow and other relevant features of population structure. The results described here and those of Slatkin & Maddison (1990) indicate that the pairwise estimate of migration rate assuming an island model of population structure, called $\hat{M}$ in both cases, is an appropriate measure of genetic similarity. Whether $\hat{M}$ is defined in terms of F_{ST} computed for pairs of populations or by Slatkin & Maddison's (1990) results for an island model depends on the data available. For allozyme data F_{ST} is appropriate and for DNA sequence data the cladistic method is appropriate. For reasons discussed above, using F_{ST}-like measures on DNA sequence data (γ_{ST} or N_{ST}) does not lead to accurate estimates of $\hat{M}$ if there is no recombination because they do not make full use of information in the data. Slatkin & Maddison (1990) illustrate how their method can be used to detect isolation by distance.

6. Conclusions

This paper has four goals. One is to show the relationship between the average coalescence times of pairs of genes and Wright's F_{ST}, which measures the extent of population subdivision. The second is to predict values of F_{ST} by computing average coalescence times in different models of subdivided populations. The third is to relate different measures of DNA sequence differences to the theory of subdivided populations. The fourth is to show that the appropriate measure of genetic similarity is the effective migration rate, $\hat{M}$, if the goal is to understand the pattern of gene flow in a subdivided population. The simplicity of the results found here follows from separating the effects of the different processes that determine the extent of geographic variation in gene frequencies.

Appendix

The problem is to find the mean coalescence time in a two-dimensional array of demes on a 'torus', that is the direct product of a circle with itself. We can do so by using the notation and formal analysis of Maruyama (1970*b*). Although this is not the standard method used in the general theory of random walks, it is quite easy and makes use of methods familiar to population geneticists. Assume that a fraction $1-m$ of the individuals in each deme remain in the deme each generation and a fraction $m/4$ go to each of the four adjacent demes. For simplicity, we will consider only the case in which $m \ll 1$ but the same method can be used when that assumption is not made. There is a total of d^2 demes in the model. Define the matrix M to be a dxd periodic matrix with elements M_{ij} ($i, j = 0, \ldots, d-1$). The non-zero elements of M are

$$M_{ii} = 1-2m, \tag{A 1a}$$
$$M_{i,i+1} = M_{i,i-1} = 2m, \tag{A 1b}$$

where the subscripts are interpreted as being evaluated modulo d. Let T be the matrix of average coalescence times of genes separated initially by i and j steps: $(T)_{ij} = \bar{t}_{ij}$. The standard theory of Markov chains (Feller, 1957) can be combined with Maruyama's (1970*b*) analysis of this population structure to write an equation that must be satisfied by T:

$$T = MTM - T_0 + U, \tag{A 2}$$

where T_0 is the matrix whose only non-zero element is $(T_0)_{00} = \bar{t}_{00}/(2N)$ and U is a matrix containing all 1s.

https://doi.org/10.1017/S0016672300029827 Published online by Cambridge University Press

Maruyama (1970*b*) showed that T can be expressed as the sum of eigenfunctions of the linear operator in (A 2):

$$T = \sum_{ij} a_{ij} E^{(i,j)}, \tag{A 3}$$

where $E^{(i,j)}$ are matrices whose klth element is $\cos(2\pi ik/d)\cos(2\pi jl/d)$, which are the eigenfunctions associated with the ijth eigenvalue, $\lambda_{ij} \approx 1 - m[2 - \cos(2\pi i/d) - \cos(2\pi j/d)]$. That is, $ME^{(i,j)}M = \lambda_{ij} E^{(i,j)}$. Note that $\lambda_{00} = 1$ and $E^{(0,0)} = U$.

The coefficients, a_{ij}, are given by

$$a_{ij} = \frac{\langle E^{(i,j)} T \rangle}{\langle E^{(i,j)} E^{(i,j)} \rangle} \tag{A 4}$$

where the inner product of two matrices is defined by

$$\langle AB \rangle = \sum_{ij} A_{ij} B_{ij}. \tag{A 5}$$

Equation (A 4) follows from (A 3) and the fact that $\langle E^{(i,j)} E^{(k,l)} \rangle = 0$ unless $i = k$ and $j = l$. It will be convenient to write

$$\langle E^{(i,j)} E^{(i,j)} \rangle = d^2 \Delta_{ij} \tag{A 5}$$

where $\Delta_{00} = 1$, $\Delta_{i0} = \Delta_{0j} = \frac{1}{2}$, and $\Delta_{ij} = \frac{1}{4} (i, j > 0)$.

By substituting (A 3) in (A 2) and then taking the inner product of the resulting equation with $E^{(i,j)}$ we obtain

$$d^2 \Delta_{ij} a_{ij} = d^2 \Delta_{ij} \lambda_{ij} a_{ij} - \frac{\bar{t}_{00}}{2N} + d^2 \delta_{i0} \delta_{j0}, \tag{A 6}$$

where $\delta_{i0} = 1$ if $i = 0$ and 0 otherwise.

If $i = j = 0$, (A 6) leaves the value of a_{00} unspecified because $\lambda_{00} = 1$ but shows immediately that $\bar{t}_{00} = 2Nd^2$, as must be true in a population with d^2 demes each of size N. For i or j not zero, (A 6) implies

$$a_{ij} = \frac{-\bar{t}_{00}}{2Nd^2 \Delta_{ij}(1 - \lambda_{ij})} = \frac{-1}{\Delta_{ij}(1 - \lambda_{ij})}. \tag{A 7}$$

To find a_{00}, we can substitute the remaining values of a_{ij} in (A 3) and use the fact that we know $\bar{t}_{00}$. The 00th element of each term in the resulting equation tells us that

$$\bar{t}_{00} = a_{00} + \sum{}' a_{ij}, \tag{A 8}$$

where $\sum'$ indicates the sum over all i and j except for $i = j = 0$. Equation (A 8) implies $a_{00} = \bar{t}_{00} - \sum' a_{ij}$ and hence

$$\bar{t}_{kl} = 2Nd^2 + \frac{1}{2m} \sum{}' \frac{1 - \cos(2\pi ik/d)\cos(2\pi jl/d)}{\Delta_{ij}\{1 - [\cos(2\pi i/d) + \cos(2\pi j/d)]/2\}}. \tag{A 9}$$

Finally, let $\sum'$ represent the sum in (A 9) so $\bar{t}_{kl} = 2Nd^2 + \sum'/2m$. The value of $F_{ST}(k, l)$ for a pair of demes k and l steps apart is, from equation (8), $(\bar{t}_{kl} - \bar{t}_{00})/(\bar{t}_{kl} + \bar{t}_{00})$. Therefore

$$F_{ST}(k, l) \approx \frac{1}{1 + 8Nmd^2/\sum'}. \tag{A 10}$$

The function $S(k.l)$ in equation (19) in the text is $\sum'/d^2$.

N. H. Barton (personal communication) has pointed out that, for large d, the $S(k, 0)$ can be approximated by

$$\frac{1}{2} \int_0^1 \int_0^1 \frac{4[1 - \cos(2k\pi x)]}{1 - [\cos(2\pi x) + \cos(2\pi y)]/2} dx\, dy \tag{A 11}$$

and that the integral with respect to y can be evaluated to obtain

$$S(k, 0) \approx \int_0^1 \frac{8[1 - \cos(2k\pi x)]}{\sqrt{[3 - 4\cos(2\pi x) + \cos^2(2\pi x)]}} dx. \tag{A 12}$$

In (A 12) d does not enter, so $S(k, 0)$ is approximately independent of d if d is large. The integral in (A 12) does not appear to be expressible in terms of elementary functions, but numerical integration shows that the approximation to the exact value $S(k, 0)$ is quite good (see Fig. 1).

This research was supported in part by Grant no. GM40282 from the U.S. National Institutes of Health. I thank N. H. Barton, R. R. Hudson, M. Lynch, T. Nagylaki and N. Takahata for helpful discussions of this topic and for helpful comments on earlier versions of this paper.

References

Crow, J. F. & Aoki, K. (1984). Group selection for a polygenic behavioral trait: estimating the degree of population subdivision. *Proceedings of the National Academy of Sciences, USA* **81**, 6073–6077.

Crow, J. F. & Kimura, M. (1970). *An Introduction to Population Genetics Theory*. New York: Harper and Row.

Ewens, W. J. (1990). Population genetics theory – the past and the future. In *Mathematical and Statistical Developments of Evolutionary Theory* (ed. S. Lessard), pp. 177–227. Amsterdam: Kluwer.

Feller, W. (1957). *An Introduction to Probability Theory and Its Applications*, 2nd edn, vol. 1. New York: Wiley.

Hey, J. (1991). A multi-dimensional coalescent process applied to multi-allelic selection models and migration models. *Theoretical Population Biology* **39**, 30–48.

Hudson, R. R. (1990). Gene genealogies and the coalescent process. In *Oxford Surveys in Evolutionary Biology* (ed. D. J. Futuyma and J. Antonovics). Oxford: Oxford University Press (in the press).

Lynch, M. & Crease, T. J. (1990). The analysis of population survey data on DNA sequence variation. *Molecular Biology and Evolution* **7**, 377–394.

Malécot, G. (1948). *Les Mathématiques de l'Hérédité*. Paris: Masson et Cie.

Maruyama, T. (1970*a*). On the rate of decrease of heterozygosity in circular stepping stone models of populations. *Theoretical Population Biology* **1**, 101–119.

Maruyama, T. (1970*b*). Effective number of alleles in a

subdivided population. *Theoretical Population Biology* **1**, 273–306.
Maruyama, T. (1971). Analysis of population structure. II. Two-dimensional stepping stone models of finite length and other geographically structured populations. *Annals of Human Genetics* **35**, 179–196.
Nei, M. (1972). Genetic distance between populations. *American Naturalist* **106**, 283–292.
Nei, M. (1973). Analysis of gene diversity in subdivided populations. *Proceedings of the National Academy of Sciences, USA* **70**, 3321–3323.
Nei, M. (1982). Evolution of human races at the gene level. In *Human Genetics, Part A: The Unfolding Genome*. (ed. B. Bohhe-Tamir, P. Cohen and R. N. Goodman), pp. 167–181. New York: Alan R. Liss.
Nei, M. (1987). *Molecular Evolutionary Genetics*. New York: Columbia University Press.
Slatkin, M. & Barton, N. H. (1989). A comparison of three indirect methods for estimating average levels of gene flow. *Evolution* **43**, 1349–1368.
Slatkin, M. & Maddison, W. P. (1989). A cladisic measure of gene flow inferred from the phylogenies of alleles. *Genetics* **123**, 603–613.
Slatkin, M. & Maddison, W. P. (1990). Detecting isolation by distance using phylogenies of genes. *Genetics* **123**, 603–613.
Strobeck, C. (1987). Average number of nucleotide differences in a sample from a single subpopulation: a test for population subdivision. *Genetics* **117**, 149–153.
Takahata, N. (1983). Gene identity and genetic differentiation of populations in the finite island model. *Genetics* **104**, 497–512.
Takahata, N. (1988). The coalescent in two partially isolated diffusion populations. *Genetical Research* **52**, 213–222.
Takahata, N. & Palumbi, S. R. (1985). Extranuclear differentiation and gene flow in a finite island model. *Genetics* **109**, 441–457.
Tavaré, S. (1984). Line-of-descent and genealogical processes and their applications in population genetics. *Theoretical Population Biology* **26**, 119–164.
Wright, S. (1922). Coefficients of inbreeding and relationship. *American Naturalist* **63**, 556–561.
Wright, S. (1951). The genetical structure of populations. *Annals of Eugenics* **15**, 323–354.
Wright, S. (1969). *Evolution and the Genetics of Population, Volume 2. The Theory of Gene Frequencies*. Chicago: University of Chicago Press.