The Average Number of Sites Separating DNA Sequences Drawn from a Subdivided Population

MONTGOMERY SLATKIN

Department of Zoology, University of California, Berkeley, California 94720

Received February 3, 1986

The "infinite sites" model in the absence of recombination is examined in a subdivided population in which there is arbitrary migration among demes. It is shown that, if the migration matrix is symmetric and irreducible, the average number of sites that differ in two alleles chosen from the same deme depends only an effective size of the whole population and not on either the elements of the migration matrix or the size of each deme separately. If there are \( n \) demes all of size \( N \), the average number of sites that differ in two alleles chosen from the same deme is \( 4nN\mu \), where \( \mu \) is the average mutation rate per site. This is the same value as for two alleles drawn from a panmictic population of size \( nN \). The average number of sites that differ in alleles drawn from the same and from different demes can provide some information about the degree of population subdivision, as is illustrated by using the data of Kreitman and Aquadé (1986, Proc. Nat. Acad. Sci. U.S.A., 83, 3562) on Drosophila melanogaster. © 1987 Academic Press, Inc.

INTRODUCTION

New DNA sequence data will pose new problems for population geneticists. One such problem arises in considering DNA sequences from individuals of the same species in different geographic locations. The number of detectable differences between alleles might reveal something about the population structure of the species from which the samples were taken. Kreitman and Aquadé (1986) present this type of data for 87 samples of a 2.7-kilobase region encompassing the ADH locus in Drosophila melanogaster. Although Kreitman and Aquadé did not determine the complete sequence of this region for each sample, they used a technique sensitive enough to detect all insertions and deletions and approximately 20% of the base pair substitutions. Of the 87 samples, 60 were from Raleigh, North Carolina, and 27 were from Putah Creek, California (near Davis).

In this note, I will investigate a model of DNA sequence evolution in a subdivided population and show that some predictions of this model are independent of the details of population subdivision and depend only on...
the total number of individuals in the population. This model is a direct extension of the model analyzed by Watterson (1975), which is Kimura's (1969) "infinite sites" model with the restriction that there is no recombination among sites. Li (1976) examined Watterson's model in the finite island model of population structure, and Griffiths (1981) examined a more general model in both the finite island and one-dimensional stepping-stone models. Griffiths' model reduces to Watterson's as a special case.

THE MODEL

Panmictic Population

We will begin by reviewing briefly the results for the infinite sites model in a single panmictic population. These results, which were first derived by Watterson (1975), will illustrate the methods used and create the context for understanding similar results for a subdivided population.

Consider a population of \( N \) randomly mating monoecious individuals and assume there are discrete, nonoverlapping generations. Assume the locus we are concerned with is made up of an infinite number of sites and that each site can mutate at a rate \( \mu \) per generation. The mutation rate per site is assumed to be sufficiently small that no site can mutate more than once. Let \( \phi_k(t) \) be the distribution of the number of sites that differ in two alleles chosen at random from the population in generation \( t \). The value \( \phi_0(t) \) is the probability of identity by descent. We can compute \( \phi_k(t+1) \) by adding over all possible outcomes of mutational events:

\[
\phi_0(t+1) = (1-\mu)^2 \left[ (1-1/2N) \phi_0(t) + 1/2N \right]
\]

and

\[
\phi_k(t+1) = (1-\mu)^2 (1-1/2N) \phi_k(t) + 2\mu(1-\mu) \phi_{k-1}(t) + \mu^2 \phi_{k-2}(t)
\]

for \( k > 0 \) with \( \phi_{-1} \equiv 0 \). Equation (1) is the usual equation for the change in the probability of identity by descent.

As noted by Watterson (1975), these equations and similar equations for the number of sites segregating in samples of more than two alleles drawn randomly from the population can be solved using probability generating functions. Let \( \Phi(s, t) = \sum s^k \phi_k(t) \) be the probability generating function (pgf) of \( \phi_k(t) \). By multiplying Eqs. (1) and (2) by \( s^k \) and adding over all values of \( k \), we obtain

\[
\Phi(s, t+1) = [(1-\mu(1-s))^2 - (1-\mu)^2/2N] \Phi(s, t) + (1-\mu)^2/2N.
\]
If $\mu \ll 1$, as is reasonable to assume, (3) can be approximated by

$$\Phi(s, t + 1) = [1 - 2\mu(1 - s) + 1/2N] \Phi(s, t) + 1/2N. \quad (4)$$

Equation (4) is the same as (1), the equation for the change in the probability of identity by descent, with the only difference being that $\mu$ is replaced by $\mu(1 - s)$. We can then use the known time-dependent and equilibrium solutions to (1) to obtain the pgf of $\phi_k(t)$. Griffiths (1981) has noted that models of subdivided populations can be simplified in the same way, and it is this feature that allows a large class of population structures to be analyzed by using the many existing results for the probabilities of identity by descent in subdivided populations.

The equilibrium solution to (4) is

$$\Phi(s) = 1/[1 + 4N\mu(1 - s)] \quad (5)$$

which implies

$$\phi_k = \theta^k/(1 + \theta)^{k + 1} \quad (6)$$

where $\theta = 4N\mu$. The number of sites that differ in two randomly chosen alleles follows a geometric distribution with parameter $\theta/(1 + \theta)$ (Watterson, 1975). The average number of sites that differ is

$$K = \Phi'(1) = \theta \quad (7a)$$

and the variance is

$$\sigma^2 = \Phi''(1) + K - K^2 = \theta(1 + \theta), \quad (7b)$$

where the prime denotes differentiation with respect to $s$. We will need these quantities later for comparison with the results for subdivided populations.

**General Stepping-Stone Model**

Consider a population subdivided into $n$ demes, with the $i$th deme being of size $N_i$. Assume that the generation consists of four steps beginning with a population of $N_i$ diploid individuals in each deme: production of an infinite gamete pool, mutation, migration of gametes, and random union of gametes to form the $N_i$ individuals in the next generation.

Let $\phi_{ii,k}$ be the probability that two randomly chosen alleles, one from deme $i$ and the other from deme $j$, differ at $k$ sites. Denoting values after mutation but before migration by a single asterisk (*), we obtain

$$\phi^*_{ii,0} = (1 - \mu)^2[(1 - 1/2N_i) \phi_{ii,0} + 1/2N_i] \quad (8a)$$

$$\phi^*_{ii,k} = (1 - \mu)^2(1 - 1/2N_i) \phi_{ii,k} + 2\mu(1 - \mu) \phi_{ii,k-1} + \mu^2 \phi_{ii,k-2} \quad (8b)$$
for $k > 0$, and for $i \neq j$

$$\phi^{*}_{i,k} = (1 - \mu)^2 \phi_{ij,k} + 2\mu(1 - \mu) \phi_{ij,k-1} + \mu^2 \phi_{ij,k-2},$$  \hspace{1cm} (8c)

where $\phi_{ij,-1} = \phi_{ij,-2} = 0$. Equations (8a) and (8b) are analogous to Eqs. (1) and (2).

For the migration stage, we define the "migration matrix" (Bodmer and Cavalli-Sforza, 1968) $M_{ij}$ as the fraction of gametes in deme $i$ that were in deme $j$ before migration. We will assume that migration is reciprocal in the sense that the migration matrix is symmetric ($M_{ij} = M_{ji}$) and that it is irreducible (which means that there are no isolated subsets of demes). Denoting values after migration by a double asterisk (**), we have

$$\phi^{**}_{i,k} = \sum M_{ii'} M_{i'j'} \phi^{*}_{i',k}$$  \hspace{1cm} (9)

where the sum is taken over only the primed variables. The random union of gametes has no effect on $\phi_{i,k}$, so $\phi^{**}_{i,k}$ is the value in the next generation.

Proceeding as in the previous section to define the pgf of $\phi_{i,k}$, $\Phi_{i}(s) = \sum s^{k} \phi_{i,k}$, we find from Eq. (8),

$$\Phi^{*}_{i}(s) = \left[ (1 - \mu(1 - s))^2 - (1 - \mu)^2/2N_i \right] \Phi_{i}(s) + (1 - \mu)^2/2N_i$$  \hspace{1cm} (10a)

$$\Phi^{*}_{i}(s) = [1 - \mu(1 - s)]^2 \Phi_{i}(s),$$  \hspace{1cm} (10b)

and from Eq. (9),

$$\Phi^{**}_{i}(s) = \sum M_{ii'} M_{i'j'} \Phi^{*}_{i'}(s).$$  \hspace{1cm} (11)

If we again assume $\mu \ll 1$, we can conveniently combine (10) and (11) to obtain a single equation for $\Phi_{i}(s)$ at equilibrium:

$$\Phi_{i}(s) = \left[ 1 - 2\mu(1 - s) \right] \sum M_{ii'} M_{i'j'} \left[ \Phi_{i'}(s) + \delta_{i',j}(1 - \Phi_{i'}(s))/2N_{i'} \right].$$  \hspace{1cm} (12)

Note that Eq. (12) is the just the equation for the equilibrium probability of identity by descent in this model of migration, with $\mu$ replaced by $\mu(1 - s)$ (cf. Cavalli-Sforza and Bodmer, 1967). It is here that we need the assumption that $M$ is irreducible. As in the case of equilibrium identities by descent in a subdivided population, there can be solutions to (12) only if that assumption is satisfied. If there were isolated subsets of demes, then in this model, $\Phi_{i}(s, t)$ would go to zero for values of $i$ and $j$ in different subsets.

The general solution to (12) is difficult to obtain, although it can in principle be found in terms of the eigenvectors of the migration matrix. It is relatively easy, however, to find the mean and variance of $k$ from (12)
without solving for $\Phi_\ell(s)$. To find $K_\ell = \Phi'_\ell(1)$, the average number of sites that differ in alleles from demes $i$ and $j$, we differentiate (12) with respect to $s$ and then set $s = 1$:

$$K_\ell = 2\mu + \sum M_\ell M_\ell (K_{\ell\ell} - \delta_{\ell\ell} K_{\ell\ell}/2N_\ell).$$

(13)

Then we sum over $i$ and $j$ and use both the fact that $\sum M_\ell = 1$ and the assumption that $M_\ell = M_{ji}$ to obtain

$$\sum (K_{\ell\ell}/N_\ell) = 4\mu n^2.$$

(14)

Equation (14) shows that, in effect, a weighted average of the average number sites that differ in two alleles drawn from the same population ($K_{\ell\ell}$) is independent of the migration matrix and depends on a single effective size of the whole population. We can see this more clearly by defining $N_a$, the harmonic mean of the sizes of each deme,

$$1/N_a = (\sum 1/N_\ell)/n$$

(15)

and the weighted average of the $K_{\ell\ell}$

$$K_0 = N_a \sum (K_{\ell\ell}/N_\ell)/n,$$

(16)

and rewriting (14) as

$$K_0 = 4nN_a\mu,$$

(17)

If $N_\ell = N$ for all $i$, then

$$K_0 = \sum K_{\ell\ell}/n = 4nN\mu,$$

(18)

in which case the average number of sites that differ in two alleles drawn from the same deme in a subdivided population depends only on the total number of individuals.

For this particular measure of allelic differences, the subdivided population is equivalent to a panmictic population of the same total size [cf. Eq. (7)]. This result is not necessarily true if the migration matrix is not symmetric and it is not true for other measures of allelic differences such as the average number of sites that differ in two alleles drawn from different demes, $K_\ell$ for $i \neq j$.

This result is consistent with the those of Li (1976) and Griffiths (1981) for the finite island model and is also consistent with that of Griffiths (1981) for the infinite linear stepping-stone model. He found that the
average number of sites that differ in two alleles drawn from any pair of populations is infinite. That is due to the assumption that the total size of the population is infinite, not, as Griffiths suggested, to particular properties of the one-dimensional population structure.

*Variation in Mutation Rates*

In the above analysis, we have assumed that the mutation rate is the same at every site. If there are differences among sites in mutation rates, the complete theory becomes much more complex. The main results of the above analysis, Eqs. (17) and (18), are still valid if we replace μ by the weighted average of the mutation rates over the sites. To see this, imagine there are distinct sets of sites each with a different mutation rate, μᵣ, for sites in set r. If we considered each set of sites separately, the average number of differences between sites in set r in alleles drawn from the same deme would be $4nN_μ_ρ$. Because this is a linear function of μᵣ, the average number of differences for all sites would be $4nN_μ$ multiplied by the weighted average of the μᵣ.

**Discussion**

The infinite sites model is formally similar to the infinite alleles model but the predictions of the two models are quite different. For the infinite alleles model, the heterozygosity within a deme, which is related to the variance in allele frequencies among demes, depends on the effective deme size (Malecot, 1969). In contrast, the average number of sites that differ in alleles from the same population depends not on the deme size but on the total population size. It is interesting and probably not accidental that Kaplan and Hudson (1987) found the same property for the average number of sites that differ in two alleles in a multigene family in which there is unbiased gene conversion. Gene conversion in their model acts like migration because it produces genetic similarities in alleles at different loci, although the two models are not formally equivalent.

The above results offer some hope that measuring the numbers of differences found in the DNA sequences from different parts of a species’ range can be useful in assessing the extent of population subdivision. Assuming that the infinite sites model is a reasonable approximation to evolution at the level of DNA, the average number of base pairs that differ between two alleles drawn from the same population $K₀$ provides an estimate of the composite parameter $4nN_μ$. The increase in the average number of sites that differ in alleles drawn from different demes depends on the migration rates and other details of the population structure.
We can use the data of Kreitman and Aquadé (1986) to see how this method can be used. Because Kreitman and Aquadé did not obtain the complete sequences of their samples, we have to estimate $K_0$ from their data. If we assume that all of the insertions and deletions but only 20% of the base pair substitutions were detected, then the data from Kreitman and Aquadé (1986, Table 1) indicate that the average number of differences between samples taken from the Raleigh population is 20.1 and from the Putah Creek population, 19.98. These numbers were obtained by adding the average numbers of insertions and deletions detected (2.1 and 1.93) to five times the number of base pair differences detected with their method (3.6 and 3.61). The infinite sites model can be applied to insertions and deletions if it is assumed that deletions are sufficiently small and infrequent that there is essentially no chance that a deletion will remove a base pair that has mutated and that insertions are sufficiently small and infrequent that a base pair substitution within an inserted segment will not occur. If either of these assumptions is not satisfied, then an important assumption of the infinite sites model, namely that mutations at different sites are independent events, cannot be true. In the data of Kreitman and Aquadé (1986), at least, insertions and deletions do appear to occur much less frequently than base pair changes, and the sizes of fragments inserted and deleted are small.

One indication of the degree of subdivision is the difference between the average number of base pair differences between alleles drawn from different and from the same population. Using the finite island model to represent the extreme in long-distance migration, the results of Li (1976) show that $K_1$, the average number of sites that differ in alleles drawn from different demes, is approximately $K_0 + (n - 1)\mu/m$, where $m$ is migration rate. For the data of Kreitman and Aquadé (1986), $K_1$ is 20.71 and $K_0 = 20.04$ so $K_1 - K_0 = 0.67$. It is impossible to estimate $n$, $m$, or $\mu$ separately, but these data suggests that either $n\mu$ is very small or $m$ is relatively large. The latter seems likely for *D. melanogaster*, a species that is commensal with humans and whose dispersal is likely to be via trucks on interstate highways.

We cannot draw any firm conclusions from the data of Kreitman and Aquadé (1986) because there are too few data and there are too many uncertainties about the biology of *D. melanogaster*, but we can see how this type of data can be used. More analysis of these models will be required to determine the statistical power and the possible limitations of this method, but the results presented here do indicate that DNA sequences will be useful for understanding population structure.
ACKNOWLEDGMENTS

This research has been supported in part by NSF Grant BSR-8500258. I thank N. H. Barton, B. C. Griffiths, M. Kreitman, W.-H. Li, N. Takahata, and two anonymous referees for helpful comments on this paper.

REFERENCES


Kimura, M. 1969. The number of heterozygous nucleotide sites maintained in a finite population due to steady flux of mutations, Genetics 61, 893–903.


