Proc. Natl. Acad. Sci. USA
Vol. 78, No. 10, pp. 6329–6333, October 1981
Genetics

Estimating genetic divergence and genetic variability with restriction endonucleases

(heterozygosity/base substitutions/population genetics/maximum likelihood)

WILLIAM R. ENGELS

Laboratory of Genetics, University of Wisconsin–Madison, Madison, Wisconsin 53706

Communicated by Harry Harris, June 4, 1981

ABSTRACT **Restriction endonucleases cut DNA at specific sites determined by the local nucleotide sequence. By comparing related DNA segments with respect to where such cuts are made, one can estimate the extent of sequence homology between the segments. Empirical methods are presented here for using these data to measure the proportion of mismatches between two sequences, the proportion of polymorphic positions in a series of sequences, or the degree of heterozygosity in a population. These methods do not require any assumptions concerning the evolutionary or population genetic processes involved. One can also use the data to calculate the precision of each of these estimates. When the positions of the cuts are not determined, these estimates can be made, using only the lengths of the resulting DNA fragments, by means of a maximum likelihood procedure. Several examples demonstrate the usefulness of these methods to study genetic differences in regions of the genome not amenable to study by other methods.**

A large part of experimental population genetics in the last two decades has dealt with protein differences within and between populations; see chapter 7 of Wright (1) for a review. More recently, the use of restriction endonucleases and accompanying DNA technology has made it possible to study differences at the level of DNA sequences. Several studies (2–5) have demonstrated variability in DNA sequences as revealed by variation in the lengths of DNA fragments after digestion by one or more sequence-specific restriction enzymes. A major advantage of this method is that it can be used for any segment of the genome, regardless of whether or not it codes for a soluble protein. It is only necessary to be able to identify the particular fragments of interest by prior purification of a given class of DNA (2, 4, 5) or by hybridization to a labeled homologous probe (3). Any variability at cleavage sites within the DNA thus identified will be detected, provided there are no large deletions or insertions causing observable length differences.

There have been several discussions of how such data may be used in genetic mapping (6), or how they may be related to models of evolutionary divergence (7–10) or steady-state population genetics (11) in order to estimate the parameters in these models. My purpose in this paper is to present empirical methods for estimating genetic divergence or population variability independent of evolutionary or population genetic models. The precision of these estimates can also be obtained with minimal assumptions concerning population structure. These methods can be applied to cases in which all restriction sites in the sample have been mapped, as well as those in which only the lengths of restriction fragments are available.

Data and Definitions. Notation will follow that of Ewens *et al.* (11) wherever possible. Our sample is assumed to consist of n homologous segments, each approximately L nucleotides long. This sample might have come from n individuals drawn from a population whose variability we wish to measure, or from representatives of n species whose genetic divergence is of interest. Alternatively, they might be homologous regions within a single genome such as the $^{A}\gamma$ and $^{G}\gamma$ human globin genes discussed below, which arose by gene duplication and have diverged in evolutionary time. Each segment in the sample is treated with one or a series of restriction endonucleases, and the lengths of the resulting fragments are determined. Assume for the moment that these lengths allow us to determine the exact point at which each enzyme cuts each fragment. (Data in which this final step was not taken, and therefore only identity of fragment lengths can be determined, will be considered below for the case of $n = 2$.) If j is the length of the recognition sequence of a given endonuclease—usually 4–6 base pairs—we may define a *site* as any sequence of j positions (base pairs). A *cleavage site* is defined as any site where at least one member of the sample was cleaved. If there are m cleavage sites, then the data consist of the values $c_1, c_2, \ldots, c_m$ ($1 \leq c_i \leq n$) representing the numbers of members in the sample cut at each site by one of the enzymes. The total number of cuts at all cleavage sites will be denoted by c.

ESTIMATORS

Frequency of polymorphism

A site may be considered polymorphic if at least one of its j positions is polymorphic in our sample. If k of the cleavage sites are observed to be polymorphic for the recognition sequence (that is, $1 \leq c_i \leq n - 1$ for exactly k sites), we might consider k/m as an estimate of the proportion of polymorphic sites in the entire segment. This estimator appears to be the one most often used (2, 3). However, as pointed out by Ewens *et al.* (11), and by Nei and Li (8) for the case of $n = 2$, this estimator contains a serious ascertainment bias because it is conditioned on the presence of at least one cut. We may correct this bias by assuming that the frequency of the recognition sequence among monomorphic sites of whatever type has the same expectation as its frequency in the sample as a whole. That is, the probability that a given site is monomorphic does not depend on its nucleotide sequence. This assumption may be written as

$$E(m - k) = (L - j + 1) \times P(\text{monomorphic site})\, P(\text{recognition sequence}), \quad [1]$$

in which $P(\ )$ indicates probability given a randomly chosen site. Noting that

$$E(c) = n(L - j + 1)\, P(\text{recognition sequence}), \quad [2]$$

we have

$$P(\text{monomorphic site}) = \frac{nE(m - k)}{E(c)}. \quad [3]$$

The publication costs of this article were defrayed in part by page charge payment. This article must therefore be hereby marked "*advertisement*" in accordance with 18 U. S. C. §1734 solely to indicate this fact.

Replacing the expectations by their observed values gives the estimator

$$\hat{P}(\text{monomorphic site}) = \frac{n(m-k)}{c}. \quad [4]$$

For a heuristic interpretation of this estimator, note that it is the proportion of all cuts in the sample that occurred in monomorphic sites. A special case of Eq. 4 was given by Nei and Li (8), (their [10] corrected by changing "x_X" to "n_X").

To estimate the polymorphism, p, at single positions, one possibility suggested by previous authors (7, 8) is to use

$$\hat{p} = 1 - \hat{P}(\text{monomorphic site})^{1/j} \quad [5]$$

from the assumption that each position is independent of its neighbors. When the experiment includes several enzymes with various lengths of recognition sequences, some weighted average of the estimates from Eq. 5 must be used (8). A reasonable alternative suggested by Ewens *et al.* (11) is to asume that a given site may be polymorphic at no more than one of its j positions. With this assumption, p may be estimated by $\hat{P}$(polymorphic site)/j, or

$$\hat{p} = \frac{c - n(m-k)}{jc}. \quad [6]$$

The interpretation of Eq. 6 is seen by noting that the numerator is the total number of cuts at or near polymorphic positions, and the denominator is the total number of positions in the sample recognized by the endonuclease. This estimator has the advantage of being easily extended to several endonucleases. The same formula may be used with c, m, and k being summed over all enzymes, and j redefined as the average length of recognition sequences weighted by the total number of cuts made by each enzyme. The interpretations of the numerator and denominator are thus preserved.

When $\hat{p}$ is used to measure genetic divergence between two genomes, it is an estimate of the proportion of base mismatches. If the n segments are taken from a larger population, $\hat{p}$ must be interpreted as an estimate of the frequency of polymorphism only in the sample at hand; it does not estimate polymorphism in the population as a whole. However, under the Wright–Fisher model of random sampling with mutation but no selection, $\hat{p}$ may be used in estimating population parameters. Thus, by the arguments of Ewens *et al.* (11), the quantity $\hat{p}/\ln n$ estimates $\theta = 4Nu$ at equilibrium, in which N is the diploid population number and u is the mutation rate.

Ewens *et al.* also showed that, under this model, the expected values of c, m, and k should be related at equilibrium in such a way that

$$p \simeq \frac{k}{2jm}, \quad [7]$$

and they suggest using this relationship to estimate p and related quantities. However, Eq. 6 might be considered preferable because it is free of assumptions concerning the population. To compare Eq. 6 with approximation 7, note that in the special case of $n = 2$, the variables c, m, and k are constrained by the identity $c = 2m - k$ irrespective of our assumptions concerning the structure of the population. Thus Eq. 6 becomes

$$\hat{p} = \frac{k}{2jm - kj}, \quad [8]$$

implying that 7 is a reasonable approximation when $k \ll m$, even though Ewens's approximation requires moderately large values of n.

Heterozygosity

If the n sequences come from a random sample of individuals in the population, we may estimate the heterozygosity in the population. Let π_i represent the true frequency in the population of the recognition sequence at site i. Then by an assumption similar to Eq. 1 we write

$$\Sigma\, \pi_i^2 = (L - j + 1)\, P(\text{homozygous site})\, P(\text{recognition sequence}),$$

which may be combined with Eq. 2 to yield

$$P(\text{homozygous site}) = \frac{n\Sigma\pi_i^2}{E(c)}.$$

Noting that

$$\Sigma\pi_i^2 = E\left[\frac{\Sigma c_i(c_i - 1)}{n(n-1)}\right]$$

from p. 69 of ref. 12, and replacing expectations by their observed values, leads to the estimator

$$\hat{P}(\text{homozygous site}) = \frac{\Sigma c_i(c_i - 1)}{c(n-1)}. \quad [9]$$

Finally, if we assume as before that a given site may be heterozygous at no more than one position so that

$$H = P(\text{heterozygous site})/j \quad [10]$$

is the heterozygosity per position, we have the estimator

$$\hat{H} = \frac{nc - \Sigma c_i^2}{jc(n-1)}. \quad [11]$$

This estimator requires no specific population genetic model. Furthermore, Eq. 10 is true under the stated assumption regardless of whether j refers to a single enzyme or to an average weighted by number of cuts over a heterogeneous series of endonucleases. This is because j measures the expected number of positions per recognized site in both cases. Therefore, the estimator 11 may be applied directly when several enzymes are used.

When $n = 2$, heterozygosity and polymorphism have the same meaning, and we should expect $\hat{H}$ to be equal to $\hat{p}$. This equality is readily demonstrated by substituting $n = 2$, $c = 2m - k$, and $\Sigma c_i^2 = 4m - 3k$ into Eqs. 6 and 11. The last substitution comes from setting $c_i = 1$ for k of the cleavage sites and $c_i = 2$ for the rest.

DNA fragment lengths

In the above discussions, I assumed that the number of cuts at each cleavage site had been determined from the resulting fragment lengths. This determination is often possible even without constructing the entire restriction map. For example, Brown (5) was able to interpret each of 59 distinct fragment length patterns as single changes (gains or losses of a recognition sequence) relative to a "typical" pattern for each enzyme without knowledge of the sequential order of the fragments. However, published data are often given as fragment lengths only (4), and the numbers of cuts can only be estimated. This situation usually arises when comparing two sequences (the case of $n = 2$), and that is the only case considered here.

Suppose F is the total number of fragments seen in both DNA segments, and G is the number of pairs of fragments with identical lengths, and thus assumed to be homologous in the two segments. The problem is to estimate m and k from the observed F and G. We may reduce this problem to the estimation of k alone by noting the relationship

$$m = (F + k + 2)/2 \quad [12]$$

Table 1. Estimates of k for linear DNA

	G												
F	0	1	2	3	4	5	6	7	8	9	10	11	12
3	5	1*											
4	6	2*											
5	7	1	1*										
6	8	2	2*										
7	9	3	1	1*									
8	10	4	2	2*									
9	11	5	3	1	1*								
10	12	6	4⁻	2	2*								
11	13	5⁺	3	3	1	1*							
12	14	6	4	2	2	2*							
13	15	7	5	3	3	1	1*						
14	16	8	6	4	2	2	2*						
15	17	9	7	5	3	3	1	1*					
16	18	10	8	6	4	2	2	2*					
17	19	11	7	5	5	3	3	1	1*				
18	20	12	8	6	4	4	2	2	2*				
19	21	13	9	7	5	5⁻	3	3	1	1*			
20	22	14	10	8⁺	6	4	4	2	2	2*			
21	23	13	11⁺	9	7	5	3	3	3	1	1*		
22	24	14	12	8	8	6	4	4	2	2	2*		
23	25	15	11	9	7	7	5	3	3	3	1	1*	
24	26	16	12	10	8	6	6	4	4	2	2	2*	
25	27	17	13	11	9	7	7	5	3	3	3	1	1*

For conditional estimates, add or subtract 2 from entries with + or − superscript; see text.
* Exact.

Table 2. Estimates of k for circular DNA

	G												
F	0	1	2	3	4	5	6	7	8	9	10	11	12
4	4												
5	5	1*											
6	6	2*											
7	7	3*	1*										
8	8	2	2*										
9	9	3	3*	1*									
10	10	4	2	2*									
11	11	5	3	3*	1*								
12	12	6	4	2	2*								
13	13	7	5	3	3*	1*							
14	14	6	4	4	2	2*							
15	15	7	5	3	3	3*	1*						
16	16	8	6	4	4	2	2*						
17	17	9	7	5	3	3	3*	1*					
18	18	10	8	6	4	2	2	2*					
19	19	11	9	7	5	3	3	3*	1*				
20	20	12	8	6	6	4	4	2	2*				
21	21	13	9	7	5	5	3	3	3*	1*			
22	22	14	10	8	6	6	4	4	2	2*			
23	23	13	11	9	7	5	5	3	3	3*	1*		
24	24	14	12	10	8	6	4	4	4	2	2*		
25	25	15	13	9	9	7	5	5	3	3	3*	1*	

* Exact.

for linear DNA because the endpoints represent cleavage sites, and

$$m = (F + k)/2 \qquad [13]$$

for circular DNA. Several authors (7, 8) have suggested ways to estimate k or related parameters from F and G by taking into account such quantities as average base composition, expected length of fragments, etc. However, I suggest using the maximum likelihood estimator, which is independent of all these quantities. The only assumption needed is that all permutations of the k dimorphic sites and the $m - k$ monomorphic sites are equally likely. Note that this assumption is implicit in all evolutionary models so far proposed (7–11), and it is independent of all parameters in these models. Because these models deal only with the frequencies of base changes, not their spatial orderings, the values of k given in Table 1 for linear DNA and in Table 2 for circular DNA may be considered maximum likelihood estimates for all these models plus any other models in which the ordering of sites is random. The tabled estimate of k and the corresponding m may be used in Eq. 8 to estimate the proportion of mismatched bases. When there are several enzymes, k and m may be estimated separately for each, then summed before applying Eq. 8.

Tables 1 and 2 were calculated as follows. Every pair of adjacent monomorphic sites results in one pair of homologous fragments, therefore

$$G = m - k - a, \qquad [14]$$

in which a is the number of runs of monomorphic sites. For example, if $k = 2$ polymorphic sites and $m - k = 5$ monomorphic site were arranged in the order *MMMMPPM*, then there would be $a = 2$ runs of *M*s and we have $G = 3$. The probability (likelihood) of a may be obtained from combinatorial considerations. Thus,

$$P(a) = \frac{\binom{m-k-1}{a-1}\binom{k+1}{a}}{\binom{m}{k}} \qquad [15]$$

for linear DNA, and

$$P(a) = \frac{\binom{m-k-1}{a-1}\binom{k}{a}}{\binom{m-1}{k-1}} \qquad [16]$$

for circular DNA. The derivations of Eqs. 15 and 16 are outlined on p. 62 of ref. 13 and p. 94 of ref. 14, respectively. Using these equations along with Eqs. 12, 13, and 14, the likelihoods of all permissible values of k were calculated up to $F = 25$, and the maximizing ks are given in Tables 1 and 2. All entries with an asterisk may be considered exact, because they represent the only values of k with nonzero likelihood. This consideration is important because published data appear to fall into the exact categories more often than not (see below), thus simplifying variance calculations.

The assumption of equiprobable permutations seems reasonable for all cases of circular DNA such as from mitochondria, but for linear DNA we might want to take into account irregularities at the termini resulting in certain permutations being forbidden. For example, one will not observe any permutation starting with *PM*... or ending with ...*MP* (*M* = monomorphic; *P* = polymorphic) because of the requirement that each segment be delimited by endonuclease cuts. We might therefore revise our assumption by stating that all permutations other than these forbidden ones are equiprobable and modify Eq. 15 accordingly. By subtracting from the numerator of Eq. 15 the number of distinguishable forbidden permutations with a runs of *M*s, and subtracting from the denominator the total number of forbidden permutations, we arrive at the required conditional probability (likelihood),

$$P(a) = \frac{\binom{m-k-1}{a-1}\left[\binom{k+1}{a} - 2\binom{k-1}{a-1} + \binom{k-3}{a-2}\right]}{\binom{m}{k} - 2\binom{m-2}{k-1} + \binom{m-4}{k-2}} \quad [17]$$

Recomputing Table 1 by using Eq. 17 resulted in only minor changes. Thus five entries were increased or decreased by 2 as indicated in Table 1 (superscript + or −, respectively). Note also that the diagonal outcomes in which $G = (F - 1)/2$ are no longer possible. Other minor irregularities at the termini such as the ambiguous sequences *MPM*... and *PPM*... might also be considered, but they occur even less frequently than the ones considered above and are likely to have less effect.

STATISTICAL PROPERTIES OF THE ESTIMATORS

The statistics $\hat{p}$ and $\hat{H}$ have no true moments because the observation $c = 0$ has positive probability, and c appears in the denominators of both expressions. However, for statistical purposes in the discussion below, I refer to the conditional distributions given $c \neq 0$. The variance estimators thus derived may be used as approximations, provided the observed c is not too close to zero. First, note that neither $\hat{p}$ nor $\hat{H}$ is unbiased, because they come from the ratio of two expectations. The mean squared error of these statistics about their true values is therefore partly from this bias, and the rest is due to sampling error. The sampling error is probably more important, and its estimation is considered in this section. Of course, the meaning of sampling variance depends on a scheme for creating hypothetical repetitions of the experiments. In the following calculations I will consider the true values of p and H to be parameters rather than variables. Thus the resulting variances are statistical sampling variances measuring the precision with which we know these parameters. They do not measure the variation in these values that would result from a hypothetical repetition of the evolutionary process that generated them.

Genetic Divergence ($n = 2$). Consider two fixed genomes for which Eq. 8 was used to estimate the proportion of mismatched bases. To repeat the experiment with the same two genomes, one could use a different set of restriction endonucleases, or move to a different part of the genome. Either way, the numbers of monomorphic and dimorphic cleavage sites can be expected to approximate two independent Poisson distributions. Thus the large-sample variance estimate is

$$V(\hat{p}) \simeq \left(\frac{\partial \hat{p}}{\partial k}\right)^2 V(k) + \left[\frac{\partial \hat{p}}{\partial (m-k)}\right]^2 V(m-k).$$

Differentiating Eq. 8 and using the observed values of k and $m - k$ as estimators of their respective variances yields

$$\begin{aligned} \hat{V}(\hat{p}) &= \frac{4km^2 - 4k^2m}{j^2c^4} \\ &= \frac{\hat{p}^2}{k}\left(1 - \frac{k^2}{c^2}\right) \\ &\simeq \hat{p}^2/k. \end{aligned} \quad [18]$$

This variance applies when k is determined directly, or when it comes from one of the asterisk-labeled entries of Tables 1 and 2.

The variance calculated by Nei and Li (8) for this situation is not a measure of the precision with which we know p for a particular pair of genomes. Instead, it is the variance that would result if "repetition" of the experiment meant returning the two genomes to their common ancestral state and allowing them to diverge again at random. To see the difference between these two variances, note that under Nei and Li's repetition scheme, we expect k and $m - k$ to be negatively correlated because the number of common ancestral restriction sites is assumed constant. However, with the scheme suggested here, the number of monomorphic, dimorphic, and noncleavage sites come from a trinomial distribution with index $L - j + 1$. Because the great majority of sites are noncleavage, the covariance between monomorphic and dimorphic cleavage sites will be negligible, and independent Poisson distributions will be approximated.

Genetic Variability. Another way to define repetition of the experiment applies when one estimates genetic variability within a population. The experiment is repeated by drawing another random sample of n homologous segments from the population and treating them with the same set of enzymes. Thus c_i will be binomially distributed with parameter π_i and index n. It is necessary to assume linkage equilibrium in the population so that these binomial distributions will be independent. Of course, this assumption can never be precisely true because each site overlaps with $2j - 2$ of its neighbors. However, for most real data, $L \gg m$ so that the probabilty of overlapping cleavage sites is negligible.

The standard large-sample approximation to the variance of p may be written as

$$\begin{aligned} V(\hat{p}) \simeq &\left[\frac{\partial \hat{p}}{\partial (m-k)}\right]^2 V(m-k) + \left(\frac{\partial \hat{p}}{\partial c}\right)^2 V(c) \\ &+ 2\left[\frac{\partial \hat{p}}{\partial (m-k)}\right]\left(\frac{\partial \hat{p}}{\partial c}\right) \mathrm{Cov}(c, m-k), \end{aligned} \quad [19]$$

in which all derivatives are evaluated at the mean values, $\bar{m}$, $\bar{c}$, and $\bar{k}$. To determine the variances and covariances, let a_i be a random variable that takes the value 1 if $c_i = n$, and 0 otherwise. Now, $m - k = \Sigma a_i$, and by our assumption of linkage equilibrium,

$$V(m-k) = \sum_i V(a_i) = \sum_i \pi_i^n (1 - \pi_i^n).$$

Similarly,

$$V(c) = \sum_i V(c_i) = \sum_i n\pi_i(1 - \pi_i)$$

and

$$\mathrm{Cov}(c, m-k) = \sum_i \mathrm{Cov}(c_i, a_i) = \sum_i n\pi_i^n (1 - \pi_i).$$

Making these substitutions and differentiating in Eq. 6 leads to

$$\begin{aligned} V(\hat{p}) \simeq \left(\frac{n}{j\bar{c}}\right)^2 \sum_i \Bigg(&\pi_i^n(1 - \pi_i^n) \\ &+ \frac{n(\bar{m} - \bar{k})\pi_i(1 - \pi_i)}{\bar{c}}\left(\frac{\bar{m} - \bar{k}}{\bar{c}} - 2\pi_i^{n-1}\right)\Bigg). \end{aligned} \quad [20]$$

Finally, the maximum likelihood estimate of $V(\hat{p})$ is obtained by replacing $\bar{m}$, $\bar{c}$, $\bar{k}$, and π_i by their respective maximum likelihood estimators, m, c, k, and c_i/n. For computational purposes, note that there are $L - j + 1$ terms in the summation in approximation **20**, but that only k of them, corresponding to the polymorphic sites, are nonzero.

The situation is somewhat more complicated when several endonucleases with various lengths of recognition sequences are involved. In this case, j must be treated as a random variable because it is an average value weighted by the numbers of cuts which are themselves random variables. Therefore approxi-

mation **20** should be expanded to include variance and covariance terms involving j. In practice, however, these terms will be negligible relative to the rest of the expression, and may be safely dropped. Thus, approximation **20** may still be used with j as the appropriate weighted average treated as a constant.

A similar procedure may be applied to the heterozygosity. Thus the large-sample approximation is

$$V(\hat{H}) \simeq \sum_i \left(\frac{\partial \hat{H}}{\partial c_i}\right)^2 V(c_i).$$

Differentiating Eq. 11 and substituting the observed value of c_i leads to the maximum likelihood estimator,

$$\hat{V}(\hat{H}) = \frac{\sum_i c_i(n - c_i)\left[2cc_i - \sum_l c_l^2\right]^2}{j^2 n(n-1)^2 c^4}. \qquad [21]$$

As in the previous case, the average length of the recognition sequence may be treated as approximately constant when there are several different endonucleases.

APPLICATIONS AND DISCUSSION

Genetic Variability. Brown (5) studied human mitochondrial DNA in 21 individuals with seven tetranucleotide restriction enzymes. From his tables 2 and 3, we see that there were $m = 244$ cleavage sites, of which $k = 45$ were polymorphic. The total number of cuts was $c = 4672$, and the frequency of polymorphism may thus be estimated at $\hat{p} = 0.0264$ from Eq. **6** with standard error 0.0015 from approximation **20**. Note that a very similar estimate of p (0.0231) is obtained from approximation **7** by the method of Ewens *et al.* (11). Under the Wright–Fisher model with neutral mutations, these estimates correspond to $\hat{\theta} = \hat{p}/\ln(21) \cong 0.0087$. It should be noted that for mitochondrial DNA, θ must be defined as Nu rather than $4Nu$ because each individual may be considered haploid, and only females reproduce.

We may also use Brown's data to estimate heterozygosity. Of course there are no true heterozygotes for mitochondrial DNA, but $\hat{H}$ is still meaningful as a measure of genetic variability. Thus, using Brown's tables 2 and 3 to obtain all c_i values for polymorphic sites, we have $\hat{H} = 0.0034$ with standard error 0.0004 from Eqs. **11** and **21**. This value is close to Nei and Li's (8) measure of "nucleotide diversity" as calculated by Brown. The latter is an unweighted average of all pairwise estimates of p, and is expected to be close to H in their model.

From comparisons between observations like these and analogous measurements for nuclear DNA, several authors (2, 8) have noted that mitochondrial DNA tends to be more variable. For example, the value of $\hat{\theta}$ calculated above for human mitochondria is approximately 35 times the estimate of Nu by Ewens *et al.* (11) from Jeffreys's (3) restriction endonuclease study of human chromosomal DNA. The only apparent exception comes from the study by Shah and Langley (15), who found that θ was approximately equivalent in *Drosophila* mitochondrial and nuclear DNA. These authors took into account the factor of 2 from haploidy, but they failed to consider the other factor of 2 resulting from cytoplasmic inheritance. With this correction, the data of Shah and Langley become qualitatively consistent with the other studies.

Estimating k from Fragment Lengths. The maximum likelihood method for estimating k may be applied to the data in figure 1 of Avise *et al.* (4), in which mitochondrial DNA from gophers produced eight fragment-length patterns when digested with the endonuclease *Hin*cII. For each of the 24 pairs of patterns, Table 2 provides an estimate of the number of dimorphic cleavage sites. For all but 10 of these comparisons, this procedure yields the exact value of k. The comparison of pattern "M" with "O" is one example in which k is not known exactly. Here $F = 14$, $G = 4$, and the only allowable values of k are 2 with likelihood 0.714, and 4 with likelihood 0.071. Therefore, the estimate is $\hat{k} = 2$ as given in Table 2, which corresponds to $\hat{p} = 0.036$ from Eqs. **8** and **13**. The alternative procedure suggested by Nei and Li (8) yields for this case $\hat{k} = 2.47$ and thus $\hat{p} = 0.044$.

Genetic Divergence. Ideal data for demonstrating the use of Eqs. **8** and **18** for estimating genetic divergence were provided by Shen *et al.* (16), who obtained the entire sequence of human $^{A}\gamma$ and $^{G}\gamma$ globin genes and the surrounding areas. This region contains a tandem duplication of approximately 5000 bases that have diverged in evolutionary time. As a thought experiment to demonstrate the method, the cleavage sites of each of 18 hexanucleotide restriction enzymes were read directly from the sequence. Any two cuts were considered homologous if their separation fell within ±50 nucleotides of the length of the duplication. The procedure yielded $m = 45$ and $k = 32$, and the estimated frequency of mismatches was $\hat{p} = 0.092$ from Eq. **8**. This estimate agrees well with the true value, 0.094, obtained directly from the sequence. The standard error from Eq. **18** was 0.016.

These data also illustrate an important defect inherent in the use of these techniques, which deal only with base substitutions. By comparing the $^{A}\gamma$ and $^{G}\gamma$ sequences, it was clear that much of the divergence was due to small insertions or deletions in addition to base substitutions (16). In extreme cases, such changes can result in sufficient alterations of fragment lengths that some homologous sites will appear to be nonhomologous. This situation would result in p and H being overestimates. However, despite this defect, the use of restriction endonucleases along with the empirical estimation procedures presented here might prove to be a useful method for studying genetic differences in regions of the genome not accessible in other ways.

I thank W. J. Ewens for providing the initial motivation for this work. Comments and suggestions came from W. J. Ewens, C. Denniston, and J. F. Crow, and unpublished work was made available by W. J. Ewens, M. Nei, and F. Tajima. This is paper number 2495 from the Laboratory of Genetics, University of Wisconsin. It was supported by Grant GM 27867 from the National Institutes of Health.

1. Wright, S. (1978) *Evolution and the Genetics of Populations* (Univ. Chicago Press, Chicago), Vol. 4.
2. Brown, W. M. & Wilson, A. C. (1979) *Proc. Natl. Acad. Sci. USA* **76**, 1967–1971.
3. Jeffreys, A. J. (1979) *Cell* **18**, 1–10.
4. Avise, J. C., Giblin-Davidson, C., Laerm, J., Patton, J. C. & Lansman, R. A. (1979) *Proc. Natl. Acad. Sci. USA* **76**, 6694–6698.
5. Brown, W. M. (1980) *Proc. Natl. Acad. Sci. USA* **77**, 3605–3609.
6. Botstein, D., White, R. L., Skolnick, M. & Davis, R. W. (1980) *Am. J. Hum. Genet.* **32**, 314–331.
7. Upholt, W. B. (1977) *Nucleic Acids Res.* **4**, 1257–1265.
8. Nei, M. & Li, W.-H. (1979) *Proc. Natl. Acad. Sci. USA* **76**, 5269–5273.
9. Kaplan, N. & Langley, C. H. (1979) *J. Mol. Evol.* **13**, 295–304.
10. Nei, M. & Tajima, F. (1981) *Genetics* **97**, 145–163.
11. Ewens, W. J., Spielman, R. S. & Harris, H. (1981) *Proc. Natl. Acad. Sci. USA* **78**, 3748–3750.
12. Kendall, M. G. & Stuart, A. (1977) *The Advanced Theory of Statistics* (Macmillan, New York), 45th Ed., Vol. 1.
13. Feller, W. (1968) *An Introduction to Probability Theory and Its Applications* (Wiley, New York), 3rd Ed., Vol. 1.
14. David, F. N. & Barton, D. E. (1963) *Combinatorial Chance* (Griffin, London).
15. Shah, D. M. & Langley, C. H. (1979) *Nature (London)* **381**, 696–699.
16. Shen, S., Slightom, J. L. & Smithies, O. (1981) *Cell*, in press.