BIOMETRICS 46, 551–582
September 1990

A BIOMETRICS INVITED PAPER WITH DISCUSSION

Sampling Strategies for Distances Between DNA Sequences

B. S. Weir and C. J. Basten

Department of Statistics, North Carolina State University,
Box 8203, Raleigh, North Carolina 27695-8203, U.S.A.

SUMMARY

An international effort is now underway to obtain the DNA sequence for the entire human genome (Watson and Jordan, 1989, *Genomics* **5**, 654–656; Barnhart, 1989, *Genomics* **5**, 657–660). This Human Genome Initiative will generate sequence data from several species other than humans, and will result in several copies per species of at least some regions of the genome. Although the project has generated much interest, it is but one aspect of the widespread effort to generate DNA sequence data. Published sequences are collected in common databases, and release 63 of GenBank in March 1990 contained 40,127,752 bases from 33,377 reported sequences (*News from GenBank* **3**; Mountain View, California: Intelligenetics, Inc., 1990). Large though this database is, it is only about 1% of the number of bases in the human genome. Interpretations of data of such magnitude are going to require the collaborative efforts of biometricians and molecular biologists, and an aim of this paper is to show that there is also a role for readers of this journal in the design of surveys of DNA sequences.

Discussion here will center on the use of sequence data in evolutionary studies, where some region of DNA is sequenced in several different species. The object is to infer the evolutionary history of that particular region, or of the species themselves. Statistical issues in the very important studies on sequences to locate and characterize regions responsible for human diseases will not be addressed here.

We will discuss appropriate ways of measuring distances between DNA sequences and of predicting the sampling properties of the distances. There are procedures for inferring evolutionary histories for a set of elements that depend on a matrix of distances between each pair of elements, and the precision of resulting trees must be influenced by the precision of the distances. We will show that account needs to be taken of two sampling processes—the sampling of sequences by the investigator ("statistical sampling"), and the sampling of genetic material involved in the formation of offspring from a parental population ("genetic sampling").

1. Sequences and Trees

The genome is the set of heriditary material transmitted from one parent to an offspring, and in the case of humans it consists of 23 chromosomes. These contain deoxyribonucleic acid (DNA) molecules which, in turn, consist of sequences of nucleotides—each being characterized by the nitrogenous base it carries. There are only four base types, *A*, *C*, *G*, *T*. Some specific regions of the genome code for proteins, the structure of the protein being determined by the order of the bases in the DNA sequence, and such regions constitute the genes. Other regions are involved in the process by which the coding information is transmitted to sites of protein synthesis, some regions are noncoding intervals between the coding units of a gene, and much of the genome has no known function. A description, and further references, of the means by which DNA sequences are determined was given by Weir (1984).

Key words: DNA sequences; Genetic distance; Genetic sampling; Evolution; Evolutionary tree; Mutation; Sampling strategy.

Evolutionary studies make use of both coding and noncoding DNA. The use of coding regions is perhaps the more obvious. A coding sequence such as that for α hemoglobin is determined for representatives of several species. The common function of hemoglobin suggests that this collection of sequences has a common ancestral source. Such sequences are said to be homologous. Although the sequences have a common function, and a common origin, they show differences resulting from mutations over time. A gene tree (Tateno, Nei, and Tajima, 1982) shows the points at which branchings occurred, leading to the present sequences, subsequent to the ancestral sequence for the group of sequences. Such a tree is shown in Figure 1. Coding sequences are thought to be constrained in the changes they may accept and still code for functional proteins. There is another class of sequences, for pseudogenes, that are noncoding. Pseudogenes are nonfunctional genes thought to have arisen from the "silencing" of duplicate genes (Nei, 1987). Such sequences are presumably not under any constraint, and so the rate at which they incorporate changes is higher than that for coding regions and more accurately reflects the rate at which mutations occur. Whether coding or noncoding regions are used, tree building depends on the idea that sequences showing a higher proportion of differences among their homologous bases are further apart in evolutionary time. It is longer since they diverged from an ancestral sequence.

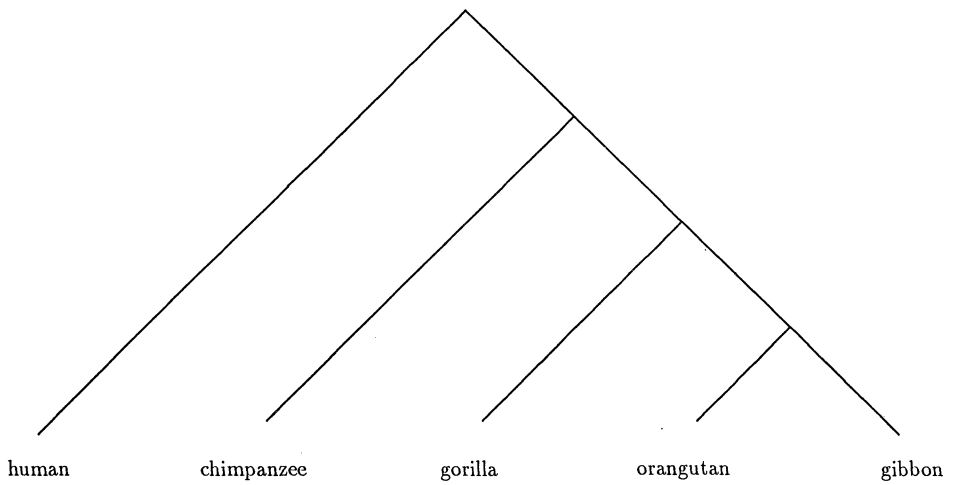

Figure 1. Gene tree for mitochondrial data for human and apes (Weir, 1989).

There are also species trees that show the evolutionary history of species (Tateno et al., 1982). Branchpoints in these trees indicate when two species became reproductively isolated from each other. In Figure 2 (Nei, 1987) we show three possible relationships between species trees and gene trees when there is polymorphism in the species. Polymorphism refers to the existence of sequence variation within a species. This can result from mutation since divergence from the last common ancestral species, or it may reflect polymorphism within that ancestral species. The (gene) tree linking a sequence from each species will depend on which sequences are drawn from the species. Information about divergence of the species themselves is contained in the species tree, and these trees can be based on distances that incorporate the variation within species. Gene trees can be constructed from distances between the individual sequences.

As yet, there has been little activity in tree building with sequences from many different regions of the genome, but we can anticipate a quickly growing use of information on

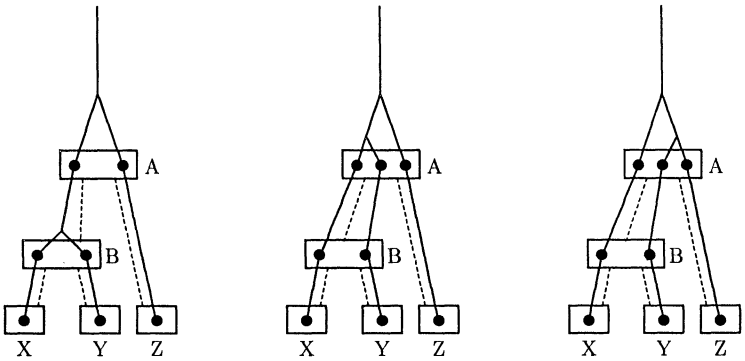

Figure 2. Possible relationships between gene and species trees [adapted from Nei (1987)]. Three extant species, X, Y, Z, have descended from ancestral species A. Species X, Y have an intermediate ancestral species B. The solid lines indicate three possible gene trees; the dotted line is the species tree.

several copies of the same sequence within one species. Such anticipation is in large part due to the recent development of techniques using the polymerase chain reaction, PCR (Wrischnik et al., 1987). This is a means of amplifying very small amounts of DNA *in vitro* to attain the amount needed for sequencing. Otherwise sequencing requires that the amount of DNA needed be produced *in vivo* by cloning—inserting the DNA in a vector such as a plasmid and letting the plasmid replicate in a bacterium to produce many copies of the inserted DNA. PCR is particularly powerful when one copy of the sequence is already available because that copy can be used to identify the same sequence using all the DNA from other individuals, in a process known as probing.

We shall show that the existence of several copies of a sequence within one species requires attention be paid to the possibility of these sequences being the same because they (recently) descended from one individual, and not merely that they failed to undergo mutation.

2. Distances Between Species

Evolutionary trees may be constructed from the information in DNA sequences by a variety of methods [see minireview by Weir (1989)], including distance matrix, parsimony, and likelihood methods. Although distance matrix methods have the disadvantage of reducing the information in two sequences to a single measure of distance between them, these methods are easy to implement and are the only ones treated here. In this paper it is assumed that sequences become more distant as a result of change over time by the process of base substitution. Other mechanisms of change, such as those causing alterations in sequence length from insertions or deletions, will not be considered.

The first such distance was introduced by Jukes and Cantor (1969). These authors worked with the sequences of amino acids constituting proteins rather than the DNA sequences of bases, but the same arguments apply. Once a pair of sequences has been correctly aligned [for review of methodology see Weir (1988a)], it is known which pairs of bases in those sequences are homologous and so have descended from a single ancestral base. If the two sequences are randomly sampled from different species, then there is a certain probability q that a pair of homologous bases are the same. This probability changes over time only because of mutation, and the simplest model supposes that mutation occurs at a rate μ per nucleotide per generation, and that each mutation is equally likely to change a base into each of the other three types. This mutation model can be expressed in terms of the

following matrix of rates from base i to base j:

	j			
i	A	C	G	T
A	—	$\mu/3$	$\mu/3$	$\mu/3$
C	$\mu/3$	—	$\mu/3$	$\mu/3$
G	$\mu/3$	$\mu/3$	—	$\mu/3$
T	$\mu/3$	$\mu/3$	$\mu/3$	—

Under this simple model only mutation can change base type at any position in a sequence, so the probability of similarity changes according to

$$q_{t+1} = (1 - \mu)^2 q_t + 2(1 - \mu)\frac{\mu}{3}(1 - q_t) + \mu^2\left[\frac{1}{3}q_t + \frac{2}{9}(1 - q_t)\right].$$

This transition equation has solution

$$q_t = \hat{q} + (q_0 - \hat{q})\left(1 - \frac{4\mu}{3}\right)^{2t},$$

where q_0 and $\hat{q} = .25$ are the initial and final values, respectively. For two sequences that have diverged for a time t since a common ancestor, $q_0 = 1$ and the quantity K defined as

$$K = \frac{3}{4}\ln\left(\frac{3}{4q_t - 1}\right)$$

has a value of

$$K \approx 2\mu t.$$

Although the mutation rate is generally unknown, K will be proportional to time since divergence and so could serve as a measure of distance between two sequences that have the same mutation rate as each other and the same rate at all positions. Of course it will apply only for finite times, before q_t has decreased to its final value of .25. The distance K can be regarded as the expected number of base substitutions during the total of $2t$ generations that have separated the two sequences (t generations along each path from a sequence to the ancestral sequence) and serves as a linearizing transformation of the quantity q. Values for q and K are plotted in Figure 3, showing behavior as a function of scaled time, $2\mu t$. Evidently the utility of q as a measure of discrimination between populations has disappeared by the time the expected number of mutations has reached $2\mu t = 3$. The linearity of K disguises this problem.

Modifications to the mutation model have been proposed, with the first being that of Kimura (1980), who allowed different rates for transitions ($A \leftrightarrow G$, $C \leftrightarrow T$) and transversions ($A, G \leftrightarrow C, T$) as in

	j			
i	A	C	G	T
A	—	β	α	β
C	β	—	β	α
G	α	β	—	β
T	β	α	β	—

An appropriate distance measure in this case is $K = (\alpha + 2\beta)t$.

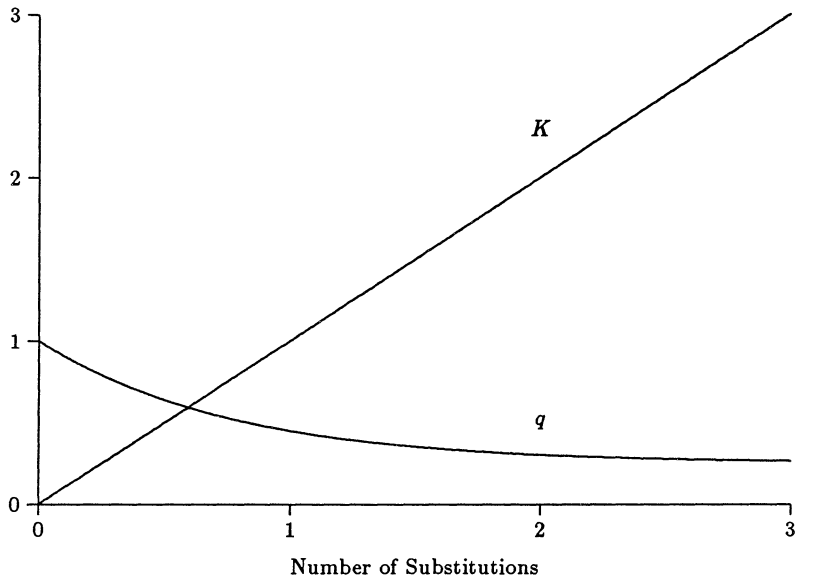

Figure 3. Behavior over time of the between-species similarity q and the Jukes–Cantor distance K. Time is measured as numbers of base substitutions, $2\mu t$.

3. Methods of Inference

When s of m nucleotide pairs in two sequences have the same base, the estimated similarity measure is

$$\tilde{q} = \frac{s}{m}.$$

We wish to determine the sampling properties of this estimator so that we can describe the properties of the sample distance

$$\tilde{K} = \frac{3}{4} \ln\left(\frac{3}{4\tilde{q} - 1}\right).$$

We employ a general methodology based on the use of indicator variables $x_{ii',l}$ defined for the two bases at position l in sequences i and i' as

$$x_{ii',l} = \begin{cases} 1, & \text{the bases are the same;} \\ 0 & \text{the bases are different.} \end{cases}$$

The number $s_{ii'}$ and the sample proportion $\tilde{q}_{ii'}$ of similar bases for the two sequences $i \neq i'$ are, therefore

$$s_{ii'} = \sum_{l=1}^{m} x_{ii',l}, \quad \tilde{q}_{ii'} = \frac{1}{m} \sum_{l=1}^{m} x_{ii',l}.$$

Taking expectations over sites from the two sequences, *and* over all pairs of sequences with the same time since divergence from the same ancestral sequence, requires the

introduction of one- and two-site ($l \neq l'$) measures of similarity, q_l and $q_{ll'}$:

$$\mathrm{E}(x_{ii',l}) = q_l,$$

$$\mathrm{E}(x_{ii',l}^2) = q_l,$$

$$\mathrm{E}(x_{ii',l}x_{ii',l'}) = q_{ll'}.$$

These quantities allow the variance of the sample proportion of similar bases to be written as

$$\mathrm{var}(\tilde{q}_{ii'}) = \frac{1}{m^2}\sum_l q_l(1 - q_l) + \frac{1}{m^2}\sum_l \sum_{l' \neq l} (q_{ll'} - q_l q_{l'})$$

Although it is usually reasonable to assume equal base similarities along short regions of a sequence ($q_l = q$, all l), it is not so clear that a common value can be assumed for the two-site similarities. We will make this assumption, however, and write the common value as q^*. Then

$$\mathrm{var}(\tilde{q}_{ii'}) = (q^* - q^2) + \frac{1}{m}(q - q^*).$$

Now there is ample empirical evidence (e.g., Nussinov, 1987) that neighboring bases in a sequence cannot be regarded as the outcomes of independent events. The frequency of any pair of bases XY in adjacent positions, for example, can differ quite markedly from the product of the frequencies of X and Y. There is also empirical evidence (Tavaré and Giddings, 1989), however, that such dependencies may not extend beyond three to four positions. For sequences of moderate length, most pairs of positions within the sequence are sufficiently far apart to be regarded as having independent similarity probabilities and these dominate the variance expression. If mutations at different sites are independent, then under the same mutation model as before, ignoring squares of μ,

$$q_{t+1}^* = \frac{4\mu}{3}q_t + \left(1 - \frac{16\mu}{3}\right)q_t^*.$$

This has solution

$$\begin{aligned} q_t^* &\approx \frac{1}{16} + \frac{6}{16}\left(1 - \frac{8\mu}{3}\right)^t + \frac{9}{16}\left(1 - \frac{16\mu}{3}\right)^t \\ &\approx (q_t)^2, \end{aligned}$$

confirming the independence of similarity probabilities at different sites. The variance of the proportion of similar sites reduces to

$$\mathrm{var}(\tilde{q}_{ii'}) = \frac{1}{m}q(1 - q).$$

The variance can be made arbitrarily small by increasing the sequence length m. This result, originally given by Jukes and Cantor (1969), shows that the similarities along a pair of sequences can be regarded as the outcomes of independent Bernoulli trials (of a mutation process). The number of similarities can be assumed to be binomially distributed

$$s \sim B(m, q).$$

Returning to the estimated distance measure, use of the delta method gives the following variance expression:

$$\begin{aligned}\operatorname{var}(\tilde{K}) &\approx \left(\frac{\partial \tilde{K}}{\partial \tilde{q}}\right)^2 \operatorname{var}(\tilde{q}) \\ &= \frac{9q(1-q)}{m(4q-1)^2}.\end{aligned}$$

4. Similarity Within Species

When more than one sequence is available for a species, it is possible to measure similarity within the species and to use this to calibrate the similarity between species. For any two sequences within a species, let Q be the probability that homologous positions have the same base. To express this in terms of the indicator variables introduced in the previous section, it is necessary to have an additional subscript j for sequences within species, and then

$$Q_l = \mathrm{E}(x_{ij,ij',l}).$$

There will generally be no need to specify the site, so that subscript l may be dropped.

With the same simple model of base substitution, in a population of N diploids Q changes over time because of mutation and drift. Two bases may be similar because of their mutation history, or because they have both descended from the same base in the previous generation:

$$Q_{t+1} \approx (1-2\mu)Q_t + \left[\frac{1}{2N} + \left(1 - \frac{1}{2N}\right)\frac{2\mu}{3}\right](1 - Q_t).$$

Writing initial and final values as Q_0 and $\hat{Q}$ leads to a solution

$$Q_t = \hat{Q} + (Q_0 - \hat{Q})\left(1 - \frac{1}{2N} - \frac{8\mu}{3}\right)^t.$$

The final value can be expressed in terms of the single quantity $\theta = 4N\mu$, provided the mutation rate is much smaller than one:

$$\hat{Q} = \frac{3+\theta}{3+4\theta}.$$

This final value is for the situation where the amount of variation being introduced by mutation balances that being lost by drift.

Nei (1972) uses the ratio of between- to within-species similarities as a measure of distance for gene frequency data. For two populations, 1 and 2, Nei's genetic identity in parametric form can be written in terms of the similarity q_{12} between species 1 and 2, and the similarities, Q_1, Q_2, within the two species:

$$I_{12} = \frac{q_{12}}{\sqrt{Q_1 Q_2}};$$

and his standard distance, D_{12}, is

$$D_{12} = -\ln I_{12}.$$

For DNA sequence data Nei (1987) employs the Jukes–Cantor distances between sequences from within or between populations. In our notation, his distance has a sample value of

$$\tilde{d} = \frac{3}{4} \ln \left[\frac{\sqrt{(4\tilde{Q}_1 - 1)(4\tilde{Q}_2 - 1)}}{4\tilde{q}_{12} - 1} \right].$$

We have considered an alternative way of comparing within- and between-species similarities to measure distance, based on concepts from analysis of variance. If n sequences are available for analysis from each of two species, there is a frequency $\tilde{p}_{il,u}$ of base type u at site l in the sample from population i. Summing over the four base types provides estimates of Q and q, as can be seen from the ANOVA format of Table 1.

Table 1
Analysis of variance format for base frequencies

Source	d.f.	Sum of squares	Expected mean square
Between	1	$\frac{2}{n} \sum_u (\tilde{p}_{1l,u} - \tilde{p}_{2l,u})^2$	$(1 - Q_l) + n(Q_l - q_l)$
Within	$2(n-1)$	$n \sum_u [\tilde{p}_{1l,u}(1 - \tilde{p}_{1l,u}) + \tilde{p}_{2l,u}(1 - \tilde{p}_{2l,u})]$	$(1 - Q_l)$

Notice that Q_l is the common expected similarity within each of the two populations. They are assumed to have been the same size since divergence. Evidently we can identify components of variance within and between populations, and we could set up a distance as the ratio of the component σ_b^2 between populations to the sum of the between- and within-population components, $\sigma_b^2 + \sigma_w^2$. We write this distance as β (Cockerham and Weir, 1987):

$$\beta = \frac{\sigma_b^2}{\sigma_b^2 + \sigma_w^2} = \frac{Q - q}{1 - q}.$$

Although both β and D increase monotonically with time, in general neither of them is directly proportional to time with the base substitution model. There does not appear to be an analogue of the quantity K when variation within species is taken into account that is proportional to time for arbitrary initial conditions. The situation is better if we make the reasonable assumption that the initial values of Q and q are the same, since both refer to the same ancestral species. If, further, we assume that the ancestral population is in equilibrium,

$$q_0 = Q_0 = \hat{Q} = \frac{3 + \theta}{3 + 4\theta},$$

we have

$$Q_t = \hat{Q},$$

$$q_t = \hat{q} + (\hat{Q} - \hat{q})\left(1 - \frac{8\mu}{3}\right)^t.$$

Nei's distance then acts as

$$D \approx \left(\frac{3}{3 + \theta}\right) 2\mu t,$$

which is directly proportional to time. This approximation holds for early times. As time becomes very large, however, D tends to the finite limit of

$$\hat{D} = -\ln\left(\frac{3 + 4\theta}{4(3 + \theta)}\right)$$

and β tends to

$$\hat{\beta} = \frac{3}{3 + 4\theta},$$

so these measures could not be used to discriminate among species that have diverged a very long time previously. There can be no discrimination among a set of species for which the equilibrium value of q has been attained. The same problem obtains when θ becomes very large, as then the range of possible values of D or β becomes very small.

In the special case of the ancestral population being in drift/mutation equilibrium, it is possible to modify the Jukes–Cantor distance to accommodate within-species variation. The distance

$$K_W = \frac{3}{4}\ln\left(\frac{Q_0 - \hat{q}}{q_t - \hat{q}}\right)$$

does change linearly with time,

$$K_W \approx 2\mu t,$$

and it has been discussed previously by Takahata (1982) and Cockerham (1984). If the ancestral polymorphism is estimated by the average extant value $(\tilde{Q}_1 + \tilde{Q}_2)/2$, setting $\hat{q}$ to $\frac{1}{4}$ gives the sample value

$$\tilde{K}_W = \frac{3}{4}\ln\left(\frac{4(\tilde{Q}_1 + \tilde{Q}_2)/2 - 1}{4\tilde{q}_{12} - 1}\right),$$

which is virtually the same as Nei's distance $\tilde{d}$.

The behavior over time of the four distances, D, β, K, and K_W, is shown in Figure 4. A typical parameter value was used in that figure: $4N\mu = .01$. In this case, the distance β loses its utility after about $10N$ generations while D remains useful for about $50N$ generations.

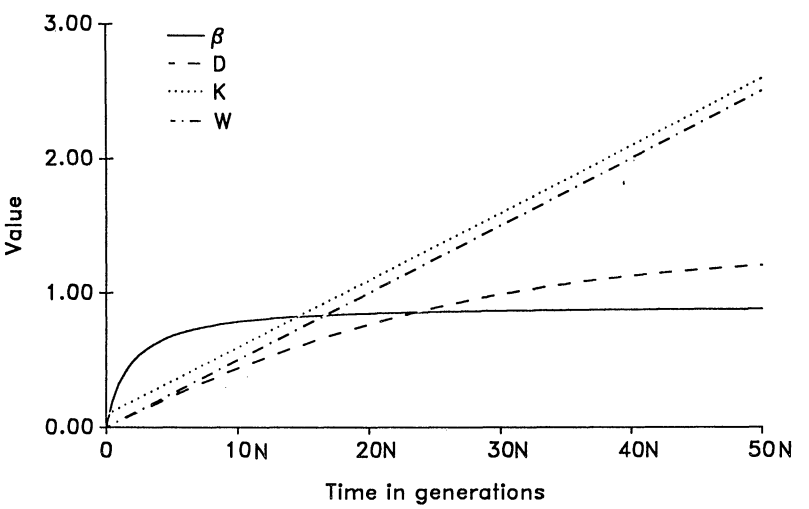

Figure 4. Behavior over time of three measures of distance between sequences that incorporate within-species variation. All curves are for $\theta = .01$.

5. Variance of Distances

This paper is concerned with sampling strategies to meet desired levels of precision for measures of distance. The object is to specify sampling parameters, the numbers and lengths of sequences, needed to meet precision levels for specific values of the genetic parameters of population size, mutation rate, and recombination rate under a specific genetic model. Because the exact constitution of a population to be sampled in a future study is not known, the sampling variances we derive account for the genetic sampling that causes replicate populations to differ even when subjected to the same forces.

5.1 *Higher-Order Similarities*

Whichever distance measure is used, there is a need to characterize the sampling properties of estimates of both Q and q as a preliminary to finding properties of the distance. Since the similarities are each defined for pairs of bases, their variances will involve up to four bases and we now introduce additional probabilities of similarity. They are shown in Figure 5, and defined in Table 2 in terms of the indicator variables. We drop the site subscript for single-site measures, and use an asterisk for measures defined at a pair of sites.

5.2 *Variances and Covariances of Similarities*

For populations i, i', observed values of the two principal similarities for any site, sample values for Q_{ii}, $q_{ii'}$, can be expressed as sums of indicator variables over the n sequences

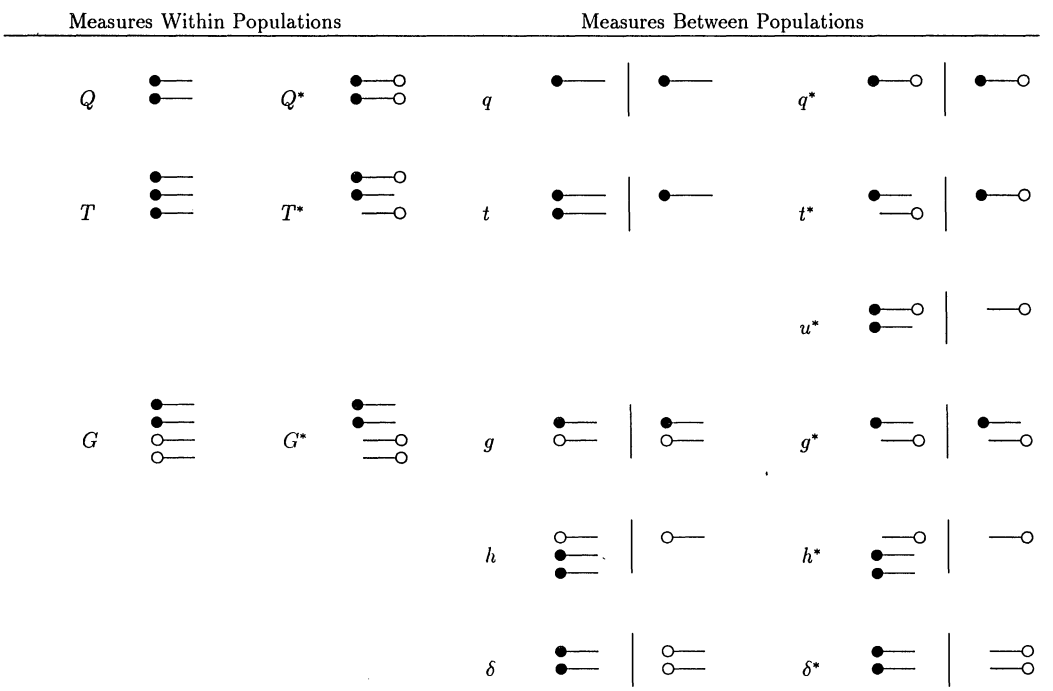

Figure 5. Similarity measures. The circles indicate the sites of interest. Two solid circles indicate that the bases are the same. Open circles are for another pair of similar sites, but they may be different from the solid circles. In the left column are the within-population measures; the right column shows the between-population measures.

Table 2
Definitions of probabilities of similarity in terms of indicator variables[a]

Number of Sequences	Number of Sites	Within populations	Between populations
2	1	$Q = \mathrm{E}(x_{ij,ij',l})$	$q = \mathrm{E}(x_{ij,i'j',l})$
	2	$Q^* = \mathrm{E}(x_{ij,ij',l}x_{ij,ij',l'})$	$q^* = \mathrm{E}(x_{ij,i'j',l}x_{ij,i'j',l'})$
3	1	$T = \mathrm{E}(x_{ij,ij',l}x_{ij,ij'',l})$	$t = \mathrm{E}(x_{ij,i'j',l}x_{ij,i'j'',l})$
	2	$T^* = \mathrm{E}(x_{ij,ij',l}x_{ij,ij'',l'})$	$t^* = \mathrm{E}(x_{ij,i'j',l}x_{ij,i'j'',l'})$
3	2		$u^* = \mathrm{E}(x_{ij,ij',l}x_{ij,i'j'',l'})$
4	1	$G = \mathrm{E}(x_{ij,ij',l}x_{ij'',ij''',l})$	$g = \mathrm{E}(x_{ij,i'j',l}x_{ij'',i'j''',l})$
	2	$G^* = \mathrm{E}(x_{ij,ij',l}x_{ij'',ij''',l'})$	$g^* = \mathrm{E}(x_{ij,i'j',l}x_{ij'',i'j''',l'})$
4	1		$h = \mathrm{E}(x_{ij,ij',l}x_{ij'',i'j''',l})$
	2		$h^* = \mathrm{E}(x_{ij,ij',l}x_{ij'',i'j''',l'})$
4	1		$\delta = \mathrm{E}(x_{ij,ij',l}x_{i'j'',i'j''',l})$
	2		$\delta^* = \mathrm{E}(x_{ij,ij',l}x_{i'j'',i'j''',l'})$

[a] The variable $x_{ij,i'j,l}$ is 1 if site l in sequence j from population i has the same base as sequence j' from population i', and is zero otherwise. In the table primes denote distinct values.

observed in each of the two populations:

$$\tilde{Q}_{il} = \frac{1}{n(n-1)} \sum_j \sum_{j' \neq j} x_{ij,ij',l};$$

$$\tilde{q}_{ii'l} = \frac{1}{n^2} \sum_j \sum_{j'} x_{ij,i'j',l}, \quad i' \neq i.$$

It is a trivial matter to allow for different numbers of sequences in the two samples. Taking the squares of each expression, and the product of the two, and then finding the expectations in terms of the measures defined in Table 2 and Figure 5, leads to the variances and the covariance of the two similarities:

$$\begin{aligned}
\mathrm{var}(\tilde{Q}_{il}) &= (G - Q^2) + \frac{4}{n}(T - G) + \frac{2}{n(n-1)}(Q - 2T + G) \\
&= \mathrm{var}(\tilde{Q}_{i'l}); \\
\mathrm{var}(\tilde{q}_{ii'l}) &= (g - q^2) + \frac{2}{n}(t - g) + \frac{1}{n^2}(q - 2t + g); \\
\mathrm{cov}(\tilde{Q}_{il}, \tilde{q}_{ii'l}) &= (h - Qq) + \frac{1}{n}(t - h). \qquad (1)
\end{aligned}$$

Even for populations that have remained distinct since divergence, the within-population similarities will be slightly correlated because of their common origin. The covariance is

$$\mathrm{cov}(\tilde{Q}_{il}, \tilde{Q}_{i'l}) = (\delta - Q^2).$$

The average within-population similarity

$$\tilde{Q}_{ii'l} = \tfrac{1}{2}(\tilde{Q}_{il} + \tilde{Q}_{i'l})$$

has variance

$$\mathrm{var}(\tilde{Q}_{ii'l}) = \left(\frac{1}{2}G + \frac{1}{2}\delta - Q^2\right) + \frac{2}{n}(T - G) + \frac{1}{n(n-1)}(Q - 2T - G) \qquad (2)$$

and covariance with the between-population measure of

$$\text{cov}(\tilde{Q}_{ii'l}, \tilde{q}_{ii'l}) = (h - Qq) + \frac{1}{n}(t - h). \tag{3}$$

Since similarities are calculated as averages over sites, we also need the covariances between sites. These are

$$\text{cov}(\tilde{Q}_{il}, \tilde{Q}_{il'}) = (G^* - Q^2) + \frac{4}{n}(T^* - G^*) + \frac{2}{n(n-1)}(Q^* - 2T^* + G^*);$$

$$\text{cov}(\tilde{q}_{ii'l}, \tilde{q}_{ii'l'}) = (g^* - q^2) + \frac{2}{n}(t^* - g^*) + \frac{1}{n^2}(q^* - 2t^* + g^*);$$

$$\text{cov}(\tilde{Q}_{il}, \tilde{q}_{ii'l'}) = (h^* - Qq) + \frac{1}{n}(u^* - h^*);$$

$$\text{cov}(\tilde{Q}_{il}, \tilde{Q}_{i'l'}) = (\delta^* - Q^2).$$

Putting all these together leads to the required variances and covariance. The lack of an l subscript on the sample similarities now indicates an averaging over m sites:

$$\begin{aligned}\text{var}(\tilde{Q}_{ii'}) = {} & \left(\frac{1}{2}\delta^* + \frac{1}{2}G^* - Q^2\right) + \frac{2}{n}(T^* - G^*) + \frac{1}{n(n-1)}(Q^* - 2T^* - G^*) \\ & + \frac{1}{2m}(\delta + G - \delta^* - G^*) + \frac{2}{nm}(T - G - T^* + G^*) \\ & + \frac{1}{n(n-1)m}(Q - 2T + G - Q^* + 2T^* - G^*);\end{aligned} \tag{4}$$

$$\begin{aligned}\text{var}(\tilde{q}_{ii'}) = {} & (g^* - q^2) + \frac{2}{n}(t^* - g^*) + \frac{1}{n^2}(q^* - 2t^* + g^*) \\ & + \frac{1}{m}(g - g^*) + \frac{2}{nm}(t - g - t^* + g^*) \\ & + \frac{1}{n^2 m}(q - 2t + g - q^* + 2t^* - g^*);\end{aligned} \tag{5}$$

$$\begin{aligned}\text{cov}(\tilde{Q}_{ii'}, \tilde{q}_{ii'}) = {} & (h^* - Qq) + \frac{1}{n}(u^* - h^*) + \frac{1}{m}(h - h^*) \\ & + \frac{1}{mn}(t - h - u^* + h^*).\end{aligned} \tag{6}$$

For a single sequence per population, $n = 1$, the within-population similarities are not calculated and the variance of the between-population similarity reduces to that given earlier. For a single site per sequence, $m = 1$, equations (4)–(6) reduce to equations (1)–(3).

Making use of the delta method for calculating variances of ratios, the approximate sampling variances of the distance measures are

$$\begin{aligned}\mathrm{var}(\tilde{K}_W) &= \frac{9}{(4Q-1)^2(4q-1)^2}[(4q-1)^2\mathrm{var}(\tilde{Q}_{ii'}) \\ &\qquad - 2(4q-1)(4Q-1)\mathrm{cov}(\tilde{Q}_{ii'}, \tilde{q}_{ii'}) + (4Q-1)^2\mathrm{var}(\tilde{q}_{ii'})] \quad (7) \\ &= \mathrm{var}(\tilde{d}),\end{aligned}$$

$$\begin{aligned}\mathrm{var}(\tilde{\beta}) = \frac{1}{(1-q)^4}[&(1-q)^2\mathrm{var}(\tilde{Q}_{ii'}) - 2(1-Q)(1-q)\mathrm{cov}(\tilde{Q}_{ii'}, \tilde{q}_{ii'}) \\ &+ (1-Q)^2\mathrm{var}(\tilde{q}_{ii'})], \quad (8)\end{aligned}$$

$$\mathrm{var}(\tilde{D}) = \frac{1}{Q^2q^2}[Q^2\mathrm{var}(\tilde{q}_{ii'}) - 2Qq\ \mathrm{cov}(\tilde{Q}_{ii'}, \tilde{q}_{ii'}) + q^2\mathrm{var}(\tilde{Q}_{ii'})], \quad (9)$$

since

$$\begin{aligned}\mathrm{var}(\tilde{Q}_{ii'}) &= \tfrac{1}{4}[\mathrm{var}(\tilde{Q}_i) + 2\ \mathrm{cov}(\tilde{Q}_i, \tilde{Q}_{i'}) + \mathrm{var}(\tilde{Q}_{i'})], \\ \mathrm{cov}(\tilde{Q}_{ii'}, \tilde{q}_{ii'}) &= \mathrm{cov}(\tilde{Q}_i, \tilde{q}_{ii'}) \\ &= \mathrm{cov}(\tilde{Q}_{i'}, \tilde{q}_{ii'}).\end{aligned}$$

Equations (7)–(9) allow the prediction of sampling variances in future populations, and hence an assessment of the effects of various combinations of sequence lengths and numbers of sequences. To allow numerical values to be found, it is necessary to evaluate the various similarity measures, and this will now be addressed.

5.3 *Transition Equations*

The similarity measures change over time under the actions of mutation, drift, and recombination. We always assume that mutation and drift are weak forces, so that second-order terms in μ or $1/N$ can be ignored. This was illustrated above for the transitions of Q and q. The remaining equations are based on the methods of Cockerham and Weir (1983) and extend those reported by Weir (1988b, 1990).

The recombination fraction r is the probability that bases at a pair of sites on one sequence have descended from bases on different sequences in that population in the previous generation. For sites from the same region of the genome, r will be small enough to ignore second-order terms involving r, μ, and $1/N$. For sites from different regions, r may be as large as .5, and then squares and products of recombination terms will need to be retained. In the former case, the forces of mutation, drift, and recombination act additively (second-order terms in r being ignored), and the transition equations may be derived easily.

Within populations at one site:

$$\begin{aligned}Q_{t+1} &= \frac{2\mu}{3} + \frac{1}{2N} + \left(1 - \frac{1}{2N} - \frac{8\mu}{3}\right)Q_t; \\ T_{t+1} &= \left(\mu + \frac{3}{2N}\right)Q_t + \left(1 - \frac{3}{2N} - 4\mu\right)T_t; \\ G_{t+1} &= \left(\frac{4\mu}{3} + \frac{1}{N}\right)Q_t + \frac{2}{N}T_t + \left(1 - \frac{3}{N} - \frac{16\mu}{3}\right)G_t.\end{aligned}$$

Within populations at two sites:

$$Q^*_{t+1} = \frac{1}{2N} + \frac{4\mu}{3}Q_t + \left(1 - \frac{1}{2N} - 2r - \frac{16\mu}{3}\right)Q^*_t + 2rT^*_t;$$

$$T^*_{t+1} = \left(\frac{1}{N} + \frac{4\mu}{3}\right)Q_t + \frac{1}{2N}Q^*_t + \left(1 - \frac{3}{2N} - r - \frac{16\mu}{3}\right)T^*_t + rG^*_t;$$

$$G^*_{t+1} = \left(\frac{4\mu}{3} + \frac{1}{N}\right)Q_t + \frac{2}{N}T^*_t + \left(1 - \frac{3}{N} - \frac{16\mu}{3}\right)G^*_t.$$

Between two populations at one site:

$$q_{t+1} = \frac{2\mu}{3} + \left(1 - \frac{8\mu}{3}\right)q_t;$$

$$t_{t+1} = \frac{\mu}{3}Q_t + \left(\frac{1}{2N} + \frac{2\mu}{3}\right)q_t + \left(1 - \frac{1}{2N} - 4\mu\right)t_t;$$

$$g_{t+1} = \frac{4\mu}{3}q_t + \frac{1}{N}t_t + \left(1 - \frac{1}{N} - \frac{16\mu}{3}\right)g_t;$$

$$h_{t+1} = \left(\frac{2\mu}{3} + \frac{1}{2N}\right)q_t + \frac{2\mu}{3}Q_t + \frac{1}{N}t_t + \left(1 - \frac{3}{2N} - \frac{16\mu}{3}\right)h_t;$$

$$\delta_{t+1} = \left(\frac{4\mu}{3} + \frac{1}{N}\right)Q_t + \left(1 - \frac{1}{N} - \frac{16\mu}{3}\right)\delta_t.$$

Between two populations at two sites:

$$q^*_{t+1} = \frac{4\mu}{3}q_t + \left(1 - 2r - \frac{16\mu}{3}\right)q^*_t + 2rt^*_t;$$

$$t^*_{t+1} = \frac{4\mu}{3}q_t + \frac{1}{2N}q^*_t + \left(1 - \frac{1}{2N} - r - \frac{16\mu}{3}\right)t^*_t + rg^*_t;$$

$$u^*_{t+1} = \frac{2\mu}{3}Q_t + \left(\frac{1}{2N} + \frac{2\mu}{3}\right)q_t + \left(1 - \frac{1}{2N} - r - \frac{16\mu}{3}\right)u^*_t + rh^*_t;$$

$$g^*_{t+1} = \frac{4\mu}{3}q_t + \frac{1}{N}t^*_t + \left(1 - \frac{1}{N} - \frac{16\mu}{3}\right)g^*_t;$$

$$h^*_{t+1} = \frac{2\mu}{3}Q_t + \left(\frac{1}{2N} + \frac{2\mu}{3}\right)q_t + \frac{1}{N}u^*_t + \left(1 - \frac{3}{2N} - \frac{16\mu}{3}\right)h^*_t;$$

$$\delta^*_{t+1} = \left(\frac{4\mu}{3} + \frac{1}{N}\right)Q_t + \left(1 - \frac{1}{N} - \frac{16\mu}{3}\right)\delta^*_t.$$

When recombination is not a very weak force, the transition equations are derived in two stages. First mutation is allowed to act, and then the joint effects of drift and recombination act on the measures after mutation. Details are given in the Appendix.

5.4 *Equilibrium and Initial Similarities*

From the transition equations for small recombination fractions, final values (denoted by carets) of the similarity probabilities can be found. They are expressed most simply in terms of $\theta = 4N\mu$ and $\Gamma = 4Nr$ and are displayed in Table 3. Compact expressions cannot be

Table 3
Equilibrium values of similarity measures for low recombination

Within populations	Between populations
$\hat{Q} = \dfrac{3+\theta}{3+4\theta}$	$\hat{q} = \frac{1}{4}$
$\hat{T} = \dfrac{(6+\theta)(3+\theta)}{(6+4\theta)(3+4\theta)}$	$\hat{t} = \hat{h} = \hat{u}^* = \hat{h}^* = \hat{Q}\hat{q}$
$\hat{G} = \dfrac{2(3+\theta)(27+12\theta+2\theta^2)}{(9+4\theta)(6+4\theta)(3+4\theta)}$	$\hat{g} = \dfrac{9+6\theta+4\theta^2}{4(3+4\theta)^2}$
	$\hat{q}^* = \hat{t}^* = \hat{g}^* = \hat{q}^2$
	$\hat{\delta} = \hat{\delta}^* = \hat{Q}^2$

found for the two-site measures within populations, although it is a simple matter to find them numerically from the following equations:

$$\begin{aligned}(3 + 3\Gamma + 8\theta)\hat{Q}^* - 3\Gamma\hat{T}^* &= 3 + 2\theta\hat{Q};\\ -6\hat{Q}^* + (18 + 3\Gamma + 16\theta)\hat{T}^* - 3\Gamma\hat{G}^* &= 4(3+\theta)\hat{Q};\\ -6\hat{T}^* + (9 + 4\theta)\hat{G}^* &= (3+\theta)\hat{Q}.\end{aligned}$$

6. Numerical Results

To investigate the temporal behavior of the distances and their sampling variances, we make the usual assumption that the ancestral population was in equilibrium for the opposing forces of drift, which decreases variation, and mutation, which increases variation. We set the within-population values equal to their equilibrium values, and the between-population initial values can all be expressed in terms of the constant within-population values:

$$\begin{aligned} q_0 &= \hat{Q}, & q_0^* &= \hat{Q}^*,\\ t_0 &= \hat{T}, & t_0^* &= u_0^* = \hat{T}^*,\\ g_0 &= h_0 = \delta_0 = \hat{G}, & g_0^* &= h_0^* = \delta_0^* = \hat{G}^*. \end{aligned}$$

We concentrate here on a single region of the DNA. There is very little recombination between adjacent sites, $r_1 \sim 10^{-8}$ (Vogel and Motulsky, 1986). At such low levels, recombination is taken to be directly proportional to physical distance. In other words, the recombination fraction r_d between sites d apart is $r_d = dr_1$. To reduce computation, we evaluate two-site measures for a region of m sites at the average recombination fraction of $(m+1)r_1/3$.

The theory presented above shows that the behavior of the similarities is determined mainly by the products $\theta = 4N\mu$ and $\Gamma = 4Nr$. We wish to investigate a "typical" natural population of size $N = 10^5 \sim 10^6$, recombination between adjacent sites of the order $r_1 \sim 10^{-8}$ and mutation also of the order $\mu \sim 10^{-8}$. These are consistent with $\theta = .01$, $\Gamma = .01$. For computational convenience, we iterated the similarity measure transition equations with $N = 10^3$ and scaled the other parameters to preserve the θ and Γ values.

Table 4
Coefficients of variation for sequence distances when $\theta = .01$ *and* $\Gamma = .01$. *At equilibrium, K and* K_W *are infinite,* $\hat{\beta} = .9868$, *and* $\hat{D} = 1.3764$.

	$n = 1$	$n = 5$			$n = 10$			$n = \infty$		
m	K	K_W	β	D	K_W	β	D	K_W	β	D
		$t = N/10$, $K = .01$, $K_W = 0$, $\beta = .05$, $D = 0$								
100 bp[a]	1.30	8.90	8.65	8.90	5.78	5.59	5.78	8.32	1.16	1.40
1 kb	.58	4.08	3.98	4.08	2.58	2.50	2.58	1.79	.59	.68
10 kb	.13	1.17	1.15	1.17	.71	.69	.71	3.05	.16	.18
		$t = N$, $K = .01$, $K_W = .01$, $\beta = .33$, $D = .01$								
100 bp	1.00	1.58	1.13	1.57	1.36	.92	1.36	2.96	.72	1.15
1 kb	.36	.64	.49	.64	.56	.40	.56	1.18	.32	.48
10 kb	.09	.16	.13	.16	.13	.10	.13	.30	.07	.11
		$t = 5N$, $K = .03$, $K_W = .03$, $\beta = .71$, $D = .03$								
100 bp	.59	.69	.27	.75	.67	.24	.73	.66	.22	.72
1 kb	.19	.24	.11	.25	.23	.10	.24	.22	.09	.24
10 kb	.05	.05	.03	.06	.05	.02	.05	.04	.02	.05
		$t = 10N$, $K = .06$, $K_W = .05$, $\beta = .83$, $D = .05$								
100 bp	.44	.48	.14	.48	.47	.12	.47	.52	.11	.46
1 kb	.14	.16	.06	.16	.16	.05	.16	.17	.05	.15
10 kb	.04	.05	.02	.05	.05	.01	.04	.05	.01	.04
		$t = 100N$, $K = .51$, $K_W = .50$, $\beta = .97$, $D = .45$								
100 bp	.19	.19	.02	.17	.19	.02	.17	.19	.01	.17
1 kb	.06	.06	.01	.05	.06	.01	.05	.06	.01	.05
10 kb	.02	.02	.00	.02	.02	.00	.02	.02	.00	.02
		$t = 1{,}000N$, $K = 5.01$, $K_W = 5.00$, $\beta = .99$, $D = 1.37$								
100 bp	6.89	6.83	.01	.12	6.82	.01	.12	6.81	.01	.12
1 kb	2.18	2.16	.00	.04	2.16	.00	.04	2.15	.00	.04
10 kb	.69	.68	.00	.01	.68	.00	.01	.68	.00	.01

[a] bp: base pair, kb: 1,000 bp.

Coefficients of variation for the three distances K, β, and D are shown in Table 4, and the same values were found with higher N values, but the same θ and Γ.

The values shown in Table 4 for the Jukes–Cantor distance K are not quite directly proportional to time since divergence. Recall that K was defined for gene trees, and applies for species trees only when the ancestral species was monomorphic. For the polymorphic ancestral case envisaged here, K changes over time according to

$$K_t = \frac{3}{4}\ln\left(\frac{3}{4q_t - 1}\right)$$

$$\approx 2\mu t + \frac{3}{4}\ln\left(\frac{3 + 4\theta}{3}\right)$$

$$= .05\left(\frac{t}{N}\right) + .0939, \quad \text{for Table 4.}$$

Although the distance was defined for the case of one sequence from each of two species, the between-sequence similarity can be defined as above when there are several sequences available. As time increases, the value of K increases without bound, and so too does its expected coefficient of variation.

The two distances β and D have substantially equal coefficients of variation, with those for β generally being lower. For both measures, the effects of increasing sequence length are much greater than those of increasing the numbers of sequences. It seems likely that evolutionary studies will be confined to cases in which no more than 10 sequences will be available per species. In this case, the coefficient of variation of distance is as low as it would be with a much larger number of sequences, provided the sequences are of length 1,000 bases (1 kb). Even then, inference is going to be difficult in early generations, $t < N$ say. On the other hand, there does not seem to be much need to increase the sequence length beyond 1 kb.

There have not yet been many empirical studies of DNA sequence variation both within and between species. The most extensive published work refers to the region containing the *Alcohol dehydrogenase* gene in species of *Drosophila*. The estimated similarities in Table 5, from Stephens and Nei (1985), refer to three species: *D. melanogaster* (11 sequences), *D. simulans*, and *D. mauritiana* (4 sequences each). Each sequence was 822 nucleotides in length. The estimated values within the three species are consistent with θ values of the order of .01, while the total map length and DNA content of *Drosophila* (Fincham, 1983) are also consistent with Γ values of .01. In other words, values shown in Table 4 may be appropriate for such studies. In particular, values shown for $t = 5N$, $n = 5$, and $m = 1$ kb are close to those needed for the data in Table 5, and illustrate that coefficients of variation of 20% are to be expected for such studies.

Table 5
Observed similarities for three Drosophila *species*

		D. melanogaster 1	*D. simulans* 2	*D. mauritiana* 3
D. melanogaster	1	$\tilde{Q}_1 = .9930$	$\tilde{q}_{12} = .9759$	$\tilde{q}_{13} = .9710$
D. simulans	2		$\tilde{Q}_2 = .9927$	$\tilde{q}_{23} = .9854$
D. mauritiana	3			$\tilde{Q}_3 = .9951$

Source: Stephens and Nei (1985).

7. Discussion

Evolutionary reconstructions can be based on measures of distance between DNA sequences. Such distances are based on the fact that mutational changes between sequences accumulate over time. The number of differences observed between two sequences needs to be modified to accommodate the chance of further mutations restoring similarities between them, and an early distance was proposed by Jukes and Cantor (1969). Their distance was appropriate for gene trees that link single sequences per species.

With information becoming available on within-species variation, it is timely to consider how variation within species may be used to calibrate that between species, and how the two sources may jointly be used in assessing evolutionary distances. Species trees can then be constructed. It may be sufficient simply to modify the Jukes–Cantor distance from K to K_W. For species in which there is little variation, Q will be close to 1 and there will be little difference between K and K_W.

The standard genetic distance D of Nei (1972) was designed to accommodate within-species variation, and it is directly proportional to divergence time for the infinite-alleles mutation model, in which every mutation results in a new form. Like K, D is then unbounded. For the mutation model considered here between four bases, Table 4 and Figure 4 indicate that D is approximately linearly related to time for early generations, maybe as many as $50N$ generations. There is a finite upper bound to D, and the utility of

the distance will decrease as time becomes large. It becomes less easy to discriminate distances between pairs of species that have diverged a very long time ago. The same statements hold for the variance component measure β, although this appears to act linearly for less than N generations. The three quantities K, d, K_W increase without bound.

The number of distinct sequences used in the construction of distances between species has relatively little effect on the predicted variance of these distances. The effects of drift impose a dependency among sequences to the point that, after a certain sample size has been reached, little information is available in further sequences. Another way of saying this is that drift leads to differences between replicate populations, and this between-population variance cannot be reduced indefinitely by further sampling within populations. For the parameter values used in Table 4, there is little point in sampling more than 10 sequences, and a case could be made for using only 5. The effects of increasing sequence length, however, are substantial. Although sampling variance cannot be eliminated (because of genetic sampling) by increasing m indefinitely, it can be made very small. The mutational process has been supposed to act independently at different sites, so that more sites provide more information.

Our treatment of the sampling variance for the quantities $\tilde{Q}$ and $\tilde{q}$ is exact, at least for small mutation rates and large population sizes. Our use of the delta method for obtaining the variances of ratios and logarithms of these quantities can only be approximate of course. As a check on our procedure, we simulated the process described here and compared predicted and simulated variances. The results are not shown here but they indicated good agreement between predicted and actual variances.

This treatment has been fairly simple, but illustrates how biometrical ideas may be of assistance in molecular evolutionary studies. There is room for extensions in the direction of greater biological realism. Nei (1987) has given a comprehensive discussion of other distance measures. He points out the need to allow for varying mutation rates along a sequence, and for distinguishing between those mutations that cause changes in encoded proteins and those that do not.

As a final comment, we stress that this paper has given a methodology for predicting sampling properties of sequence statistics. Questions of estimating variances from single data sets have not been addressed.

ACKNOWLEDGEMENTS

Helpful comments on a draft of this paper were made by Drs Clark Cockerham, Robert Curnow, and Avigdor Beiles. We appreciate the accompanying commentaries by Drs Jonathan Arnold, Norman Kaplan, and Simon Tavaré. We are in substantial agreement with these three authors.

This is Paper Number 12490 of the Journal Series of the North Carolina Agricultural Experiment Station, Raleigh, North Carolina 27695-7643, U.S.A. This investigation was supported in part by NIH Research Grant GM 32518, and by a Sloan Foundation Postdoctoral Fellowship to CJB. Computing was supported in part by a grant from the Pittsburgh Supercomputing Center through the NIH Division of Research Resources cooperative agreement U41 RR04154. Part of this material was presented at the XIV International Biometric Conference, Namur, Belgium, in 1988.

RÉSUMÉ

Un important effort international a été entrepris pour obtenir la séquence d'ADN du génome humain dans son intégralité (Watson et Jordan, 1989, *Genomics* **5**, 654–656; Barnhart, 1989, *Genomics* **5**, 657–660). Cette "Initiative pour le génome Humain" fournira des données sur les séquences de beaucoup d'autres espèces que l'espèce humaine, et l'on pourra disposer, au moins pour certaines régions du génome, de nombreuses copies par espèces. Bien que ce projet ait suscité un grand intérêt,

il ne s'agit là que de l'un des aspects de l'effort entrepris de par le monde pour obtenir des séquences d'ADN. Les séquences publiées sont réunies dans des bases de données, et la version n° 63 de "GenBank" contenait en mars 1990 40,127,752 bases pour 33,377 séquences répertoriées (*News from GenBank*, Vol. 3, n° 1; Mountain View, California: Intelligenetics). Malgré la taille de cette banque de données, elle ne contient qu'un pour cent environ de la quantité totale de bases du génome humain! L'interprétation d'une telle quantité de données exigera la collaboration assidue des biométriciens et des biologistes moléculaires, et l'un des objectifs de cet article est également de montrer aux lecteurs de cette revue qu'ils ont eux-mêmes un rôle à jouer dans la planification des études sur les séquences de l'ADN.

Nous focaliserons notre discussion sur l'utilisation de ces données dans les études portant sur l'évolution, lorsqu'une même région de l'ADN est séquencée chez plusieurs espèces différentes. L'objectif est d'analyser l'évolution dans le temps de cette région particulière, ou de comprendre l'évolution des espèces elle-mêmes. Nous ne nous intéresserons pas ici aux considérations statistiques portant sur les études de séquences ayant pour objectif de localiser et de caractériser les régions responsables de maladies humaines.

Nous discuterons les méthodes adaptées à la mesure de distances entre séquences d'ADN et à précision de distribution d'enchantillonnage de ces distances. Pour déduire l'histoire de l'évolution d'un ensemble d'éléments, il existe des procédures dépendant de la matrice des distances entre paires d'éléments. La précision des arbres qui en résultent est fonction de la précision des distances. Nous montrerons qu'il faut considérer avec attention les deux procédures d'echantillonnage: l'échantillonnage des séquences par l'expérimentateur ("échantillonnage statisique") et l'échantillonnage du matériel génétique dans la descendance d'une population ("échantillonnage génétique").

References

Barnhart, B. J. (1989). The Department of Energy (DOE) Human Genome Initiative. *Genomics* **5,** 657–660.

Cockerham, C. C. (1984). Drift and mutation with a finite number of allelic states. *Proceedings of the National Academy of Sciences USA* **81,** 530–534.

Cockerham, C. C. and Weir, B. S. (1983). Variance of actual inbreeding. *Theoretical Population Biology* **23,** 85–109.

Cockerham, C. C. and Weir, B. S. (1987). Correlations, descent measures: Drift with migration and mutation. *Proceedings of the National Academy of Sciences USA* **84,** 8512–8514.

Fincham, J. R. S. (1983). *Genetics*. Bristol: Wright.

Intelligenetics, Inc. (1990). *News from GenBank*, Volume 3, Number 1. Mountain View, California: Intelligenetics.

Jukes, T. H. and Cantor, C. R. (1969). Evolution in protein molecules. In *Mammalian Protein Metabolism*, H. N. Munro (ed.), 21–123. New York: Academic Press.

Kimura, M. (1980). A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. *Journal of Molecular Evolution* **16,** 111–120.

Nei, M. (1972). Genetic distance between populations. *American Naturalist* **106,** 283–292.

Nei, M. (1987). *Molecular Evolutionary Genetics*. New York: Columbia University Press.

Nussinov, R. (1987). Nucleotide quartets in the vicinity of eukaryotic transcriptional initiation sites: Some DNA and chromatin structural implications. *DNA* **6,** 613–622.

Stephens, J. C. and Nei, M. (1985). Phylogenetic analysis of polymorphic DNA sequences at the *Adh* locus in *Drosophila melanogaster* and its sibling species. *Journal of Molecular Evolution* **22,** 289–300.

Takahata, N. (1982). Linkage disequilibrium, genetic distance and evolutionary distance under a general model of linked genes or a part of the genome. *Genetical Research* **39,** 63–77.

Tateno, Y., Nei, M., and Tajima, F. (1982). Accuracy of estimated phylogenetic trees from molecular data. I. Distantly related species. *Journal of Molecular Evolution* **18,** 387–404.

Tavaré, S. and Giddings, B. W. (1989). Some statistical aspects of the primary structure of nucleotide sequences. In *Mathematical Methods for DNA Sequences*, M. S. Waterman (ed.), 117–132. Boca Raton, Florida: CRC Press.

Vogel, F. and Motulsky, A. G. (1986). *Human Genetics*, 2nd edition. Berlin: Springer-Verlag.

Watson, J. D. and Jordan, E. (1989). The Human Genome Project at the National Institutes of Health. *Genomics* **5,** 654–656.

Weir, B. S. (1984). Statistical analysis of molecular genetic data. *IMA Journal of Mathematics Applied in Medicine and Biology* **2,** 1–39.

Weir, B. S. (1988a). Statistical analysis of DNA sequences. *Journal of the National Cancer Institute* **80,** 395–406.

Weir, B. S. (1988b). Sampling properties of sequence statistics. In *Proceedings of the* XIV *International Biometric Conference*, 145–156. Alexandria, Virginia: The Biometric Society.
Weir, B. S. (1989). Building trees with DNA sequences. *Biometric Bulletin* **6,** 21–23.
Weir, B. S. (1990). Variation in sequence distances. In *Molecular Evolution*, M. T. Clegg and S. J. O'Brien (eds), 281–288. New York: Alan R. Liss.
Wrischnik, L. A., Higuchi, R. G., Stoneking, M., Erlich, H. A., Arnheim, N., and Wilson, A. C. (1987). Length mutations in human mitochondrial DNA: Direct sequencing of enzymatically amplified DNA. *Nucleic Acids Research* **15,** 529–542.

Received November 1989; *revised January* 1990; *accepted January* 1990.

APPENDIX

A.1 *Mutation Transition Equations*

If an m subscript indicates the results of mutation, the mutation-only transitions are as follows.
Within populations at one site:

$$Q_{t,m} = \frac{2\mu}{3} + \left(1 - \frac{8\mu}{3}\right)Q_t;$$

$$T_{t,m} = \mu Q_t + (1 - 4\mu)T_t;$$

$$G_{t,m} = \frac{4\mu}{3}Q_t + \left(1 - \frac{16\mu}{3}\right)G_t.$$

Within populations at two sites:

$$Q^*_{t,m} = \frac{4\mu}{3}Q_t + \left(1 - \frac{16\mu}{3}\right)Q^*_t;$$

$$T^*_{t,m} = \frac{4\mu}{3}Q_t + \left(1 - \frac{16\mu}{3}\right)T^*_t;$$

$$G^*_{t,m} = \frac{4\mu}{3}Q_t + \left(1 - \frac{16\mu}{3}\right)G^*_t.$$

Between populations at one site:

$$q_{t,m} = \frac{2\mu}{3} + \left(1 - \frac{8\mu}{3}\right)q_t;$$

$$t_{t,m} = \frac{\mu}{3}Q_t + \frac{2\mu}{3}q_t + (1 - 4\mu)t_t;$$

$$g_{t,m} = \frac{4\mu}{3}q_t + \left(1 - \frac{16\mu}{3}\right)g_t;$$

$$h_{t,m} = \frac{2\mu}{3}(Q_t + q_t) + \left(1 - \frac{16\mu}{3}\right)h_t;$$

$$\delta_{t,m} = \frac{4\mu}{3}Q_t + \left(1 - \frac{16\mu}{3}\right)\delta_t.$$

Between populations at two sites:

$$q^*_{t,m} = \frac{4\mu}{3}q_t + \left(1 - \frac{16\mu}{3}\right)q^*_t;$$

$$t^*_{t,m} = \frac{4\mu}{3}q_t + \left(1 - \frac{16\mu}{3}\right)t^*_t;$$

$$u^*_{t,m} = \frac{2\mu}{3}(Q_t + q_t) + \left(1 - \frac{16\mu}{3}\right)u^*_t;$$

$$g^*_{t,m} = \frac{4\mu}{3}q_t + \left(1 - \frac{16\mu}{3}\right)g^*_t;$$

$$h^*_{t,m} = \frac{2\mu}{3}(Q_t - q_t) + \left(1 - \frac{16\mu}{3}\right)h^*_t;$$

$$\delta^*_{t,m} = \frac{4\mu}{3}Q_t + \left(1 - \frac{16\mu}{3}\right)\delta^*_t.$$

A.2 *Drift and Recombination Transition Equations*

The drift and recombination transitions make use of the probabilities of obtaining sequences within a population from a specified number of parental sequences in the previous generation. For diploid populations of size N, i.e., $2N$ copies of each sequence in each generation, the assumptions of random mating, including a random amount of selfing, lead to the following expressions for these probabilities. Two sequences descend from one or two parental sequences with probabilities P^2 and P^{11}, where

$$P^2 = \frac{1}{2N}, \quad P^{11} = \frac{2N-1}{2N}.$$

Three sequences descend from one, two, or three sequences with probabilities P^3, P^{21}, and P^{111}, where

$$P^3 = \frac{1}{4N^2}, \quad P^{21} = \frac{3(2N-1)}{4N^2}, \quad P^{111} = \frac{(2N-1)(2N-2)}{4N^2}.$$

Four sequences descend from one, three, or four sequences with probabilities P^4, P^{211}, and P^{1111}. They descend two from each of two sequences with probability P^{22}, and three from one sequence and the fourth from another with probability P^{31}. These quantities are

$$P^4 = \frac{1}{8N^3}, \quad P^{211} = \frac{6(2N-1)(2N-2)}{8N^3}, \quad P^{1111} = \frac{(2N-1)(2N-2)(2N-3)}{8N^3},$$

$$P^{31} = \frac{4(2N-1)}{8N^3}, \quad P^{22} = \frac{3(2N-1)}{8N^3}.$$

With these probabilities, the transitions equations are as follows.

Within populations at one site:

$$Q_{t+1} = P^2 + P^{11}Q_{t,m};$$

$$T_{t+1} = P^3 + P^{21}Q_{t,m} + P^{111}T_{t,m};$$

$$G_{t+1} = P^4 + \tfrac{1}{6}P^{211}(2Q_{t,m} + 4T^*_{t,m}) + P^{1111}G_{t,m} + P^{31}Q_{t,m} + \tfrac{1}{3}P^{22}(1 + 2Q_{t,m}).$$

Within populations at two sites:

$$Q^*_{t+1} = P^2[(1-r)^2 + 2r(1-r)(2Q_{t,m} - T^*_{t,m}) + r^2(1 + Q^*_{t,m} - G^*_{t,m})]$$
$$+ P^{11}[(1-r)^2 Q^*_{t,m} + 2r(1-r)T^*_{t,m} + r^2 G^*_{t,m}];$$
$$T^*_{t+1} = P^3[(1-r) + rQ_{t,m}] + \tfrac{1}{3}P^{21}[2Q_{t,m} + (1-r)Q^*_{t,m} + rT^*_{t,m}]$$
$$+ P^{111}[(1-r)T^*_{t,m} + rG^*_{t,m}];$$
$$G^*_{t+1} = P^4 + \tfrac{1}{6}P^{211}(Q_{t,m} + \tfrac{4}{6}T^*_{t,m}) + P^{1111}G^*_{t,m}$$
$$+ P^{31}Q_{t,m} + \tfrac{1}{3}P^{22}(1 + 2Q^*_{t,m}).$$

Between populations at one site:

$$q_{t+1} = q_{t,m};$$
$$t_{t+1} = P^2 q_{t,m} + P^{11} t_{t,m};$$
$$g_{t+1} = (P^2)^2 q_{t,m} + 2P^2 P^{11} t_{t,m} + (P^{11})^2 g_{t,m};$$
$$h_{t+1} = P^3 q_{t,m} + \tfrac{1}{3}P^{21}(q_{t,m} + 2t_{t,m}) + P^{111}h_{t,m};$$
$$\delta_{t+1} = (P^2)^2 + 2P^2 P^{11} Q_{t,m} + (P^{11})^2 \delta_{t,m}.$$

Between populations at two sites:

$$q^*_{t+1} = (1-r)^2 q^*_{t,m} + 2r(1-r)t^*_{t,m} + r^2 g^*_{t,m};$$
$$t^*_{t+1} = P^2[(1-r)q^*_{t,m} + rg^*_{t,m}] + P^{11}[(1-r)t^*_{t,m} + rg^*_{t,m}];$$
$$u^*_{t+1} = P^2 q_{t,m} + P^{11}[(1-r)u^*_{t,m} + rh^*_{t,m}];$$
$$g^*_{t+1} = (P^2)^2 q^*_{t,m} + 2P^2 P^{11} t^*_{t,m} + (P^{11})^2 g^*_{t,m};$$
$$h^*_{t+1} = P^3 q_{t,m} + \tfrac{1}{3}P^{21}(q_{t,m} + 2u^*_{t,m}) + P^{111}h^*_{t,m};$$
$$\delta^*_{t+1} = (P^2)^2 + 2P^2 P^{11} Q_{t,m} + (P^{11})^2 \delta^*_{t,m}.$$

DISCUSSION ON THE PAPER BY B. S. WEIR AND C. J. BASTEN

J. Arnold (Department of Genetics, Biological Science Building, The University of Georgia, Athens, Georgia 30602, U.S.A.)

The similarity (descent) measures, which Weir and Basten introduce to study sequence variation, fall into two classes. One class of measures (Q, T, Q^*, T^*, etc.) decay to their equilibrium values extremely quickly (on the order of $1/N$ for a decay rate) and are associated with the within-species component of sequence variation. These measures will be extremely useful in distinguishing historical processes such as drift and mutation from the varying kinds of natural selection operating on sequence variation at the molecular level. The second class of measures (q, t, q^*, t^*, etc.), on the other hand, decay to their equilibrium values extremely slowly (on the order of μ for a decay rate) and are associated with the between-species component of sequence variation. These measures will be extremely useful in phylogenetic tree reconstruction. For many cases, the time scales for the dynamics of these two classes of measures are nonoverlapping, although there are some interesting cases such as hybrid zones in which both suites of measures need to be considered simultaneously.

As an example of the difference in time scales, iterating the recursions for the similarity measures between two species initially identical for $N = 1{,}000$, $\mu = r = 10^{-8}$, and $n = 5$ is instructive. These parameter choices might be typical of mitochondrial RFLPs in a mouse population. The equilibrium value for Q is achieved almost immediately ($4N$) but q reaches its equilibrium of $\frac{1}{4}$ only after $190{,}000N$ generations. While Q will be extremely useful in identifying selective constraints on genes in mtDNA within the lifetime of species, the measure q can be very useful in reconstructing a phylogeny of closely related species. Some of the clock-like properties of the other between-species measures need to be examined in addition to monotone transformations of q for phylogenetic tree reconstruction (i.e., q^* and t^*).

One of the limitations of the Weir and Basten formulation is that it oversimplifies how natural selection operates on sequences. Treating the state space in the mutation model as the possible nucleotides A, G, C, and T is inadequate for most empirical studies. One implication of the Jukes–Cantor model is that the steady-state frequencies of A, G, C, and T under mutation should all be $\frac{1}{4}$, which is incorrect for virtually all organisms with the exception of *E. coli*. Each species has a characteristic pattern of codon usage, which is not predicted from the Jukes–Cantor mutation model. Substitution rates vary with codon context (silent vs replacement), with position (active site vs side chain) along a sequence, and by codon (tRNA abundance). Minimally, the state space in the mutation model needs to be all 256 tetranucleotides (Arnold et al., 1988), if we are to explain the oligonucleotide composition of genomes by a mutation/recombination/drift balance with some purifying selection. Within this richer state space, a realistic mutation model could be formulated to study the similarity measures.

Lastly, while the focus of this paper is on the use of similarity measures in phylogenetic reconstruction, the between-species measures involving multiple sites will prove extremely useful in genomic mapping. For example, with pulsed field electrophoresis, extremely large fragments of DNA can be run out on a gel with the potential for picking up more than one RFLP per fragment. These so-called class II polymorphisms (Meagher, McLean, and Arnold, 1988) can be utilized to estimate how many bases there are in a centimorgan. This information is necessary in exploiting the RFLP map to walk to an interesting locus. Breeders can routinely generate class II RFLPs from interspecies crosses in plants, provided the species are sufficiently diverged. The variable $(1 - x_{ij,i'j',l})(1 - x_{ij,i'j',l'})$ indicates a class II RFLP. The expectation of this indicator variable in terms of the similarity measures is $1 - 2q + q^*$. For the mouse population above, we will have to wait $16{,}000N$ generations before the chance of a class II polymorphism (7%) is sufficiently high to be useful. Utilizing the first estimate of silent substitution rate for plant nuclear DNA (Meagher, Berry-Lowe, and Rice, 1989) of $\mu = 6.6 \times 10^{-9}$ and leaving the remaining parameters unchanged in the mouse example, we will need to wait $25{,}000N$ generations for the appearance of a usable species frequency of class II polymorphisms.

References

Arnold, J., Cuticchia, A. J., Newsome, D. A., Jennings, W. W., and Ivarie, R. (1988). Mono- through hexanucleotide composition of the sense strand of yeast DNA: A Markov chain analysis. *Nucleic Acids Research* **16,** 7145–7158.

Meagher, R. B., Berry-Lowe, S., and Rice, K. (1989). Molecular evolution of the small subunit of ribulose bisphosphate carboxylase: Nucleotide substitution and gene conversion. *Genetics* **123,** 845–863.

Meagher, R. B., McLean, M. D., and Arnold, J. (1988). Recombination within a subclass of restriction fragment length polymorphisms may help link classical and molecular genetics. *Genetics* **120,** 809–818.

Norman Kaplan (National Institute of Environmental Health Sciences, P.O. Box 12233, Research Triangle Park, North Carolina 27709, U.S.A.)

I want to elaborate on the theoretical model underlying much of the statistical analysis in the paper by Weir and Basten. In particular, I hope to shed more light on the role of (a) recombination, (b) the model describing the mutational process for an individual base, and (c) intrapopulation sequence data. In what follows I will be as general as possible.

Suppose we randomly sample from each of two closely related populations, a DNA sequence in a region of the genome that contains m nucleotides. As a result of "genetical sampling," the two sequences have a "genealogical history" (Hudson, 1983). This history consists of m ancestral trees, one for each nucleotide. The ancestral tree traces the genealogy of the sample back to the most recent common ancestor of the nucleotide. For two sequences the ancestral tree of a nucleotide is particularly simple and is described in Figure 1.

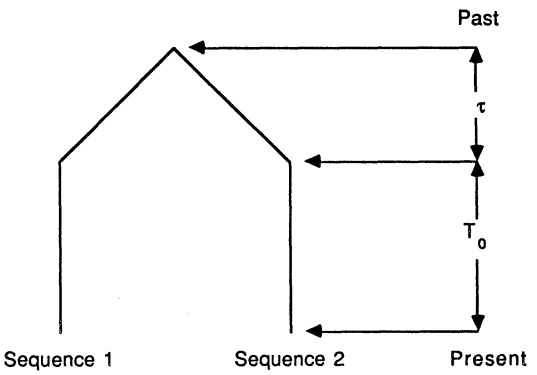

Figure 1. An ancestral tree.

The T_0 ancestral generation is when the two populations split, and $T_0 + \tau$ is the ancestral generation when the common ancestor of the two sequences occurred. The ancestral tree is important because only mutations occurring since the time of the common ancestor of the two sequences can segregate in the sample. T_0 is the same for all m bases, but τ is random and can vary from nucleotide to nucleotide, depending on the rate of recombination for the region. To allow for this we let τ_l denote the value of τ for the lth nucleotide, $1 \leq l \leq m$. The $\{\tau_l\}$ are identically distributed and their common distribution depends on the population genetics model that governs the evolution of the population. If the region is completely linked, then the ancestral tree is the same for all m bases and all the $\{\tau_l\}$ are equal. As the rate of recombination increases, the correlation between the $\{\tau_l\}$ decreases (Hudson, 1983).

Variation is assumed to be selectively neutral in the sense that no mutation affects the genetical sampling. Furthermore, it is assumed that the mutational process is the same for each nucleotide, and that the probability of a particular nucleotide undergoing mutation per chromosome, per generation depends only on the state of the nucleotide and is independent of all else. For $1 \leq l \leq m$, let

$$y_l = \begin{cases} 1 & \text{if the } l\text{th base in the two sequences is different,} \\ 0 & \text{if the } l\text{th base is the same.} \end{cases}$$

It follows from the assumptions that conditional on the genealogical history of the two sequences, the $\{y_l,\ 1 \leqslant l \leqslant m\}$ are independent random variables and

$$\mathrm{E}(y_l \mid T_0 + \tau_l) = 1 - \sum_{j=1}^{4} \sum_{k=1}^{4} p_j(T_0 + \tau_l) P_{jk}^2(T_0 + \tau_l), \quad 1 \leqslant l \leqslant m,$$

where $\{p_j(T_0 + \tau_l),\ 1 \leqslant j \leqslant 4\}$ are the population frequencies of the four bases in the $T_0 + \tau_l$ ancestral generation and $P_{jk}(T_0 + \tau_l)$ is the probability that the lth base is currently in state k, given it was in state j in the $T_0 + \tau_l$ ancestral generation, $1 \leqslant j, k \leqslant 4$.

It is commonly assumed that for $1 \leqslant j, k \leqslant 4$, and $t > 0$,

$$P_{jk}(t) = P(X(t) = k \mid X(0) = j),$$

where X is a four-state Markov chain. Also it is assumed that the ancestral population is in equilibrium, and so

$$p_j(T_0 + \tau_l) = p_j, \quad 1 \leqslant j \leqslant 4,$$

where $\{p_j\}$ are the stationary probabilities associated with the Markov chain X. The various models in the literature (e.g., Jukes and Cantor, 1969; Kimura, 1980, 1981; Takahata and Kimura, 1981; Gojobori, Ishii, and Nei, 1982) differ only in their assumptions about the parameters of the X process.

Since the two populations are closely related, it is assumed that mutation is rare. Thus, if λ_j is the rate of mutation per genome per generation for a nucleotide in state j, then

$$\begin{aligned}\mathrm{E}(y_l \mid T_0 + \tau_l) &\approx 1 - \sum_{j=1}^{4} p_j (1 - e^{-\lambda_j (T_0 + \tau_l)})^2 \\ &\approx 2\left(\sum_{j=1}^{4} p_j \lambda_j\right)(T_0 + \tau_l).\end{aligned}$$

For some of the models (e.g., Jukes and Cantor, 1969), explicit expressions can be obtained for $\mathrm{E}(y_l \mid T_0 + \tau_l)$, but this requires specific assumptions about the X process. My goal here is to keep the discussion general.

The quantity

$$D = 2\left(\sum_{j=1}^{4} p_j \lambda_j\right) T_0$$

is one measure of distance that has been considered (Nei, 1987). D is interpreted as the expected number of mutations per nucleotide between two random DNA sequences obtained from different populations that diverged T_0 generations in the past. For the Jukes and Cantor model, $p_j = \frac{1}{4}$ and $\lambda_j = \mu$ $(1 \leqslant j \leqslant 4)$, and so $D = 2\mu T_0$.

We next consider how to estimate D. Let

$$\begin{aligned}S &= \text{average number of segregating} \\ &\qquad \text{bases between two sequences} \\ &= \frac{1}{m} \sum_{l=1}^{m} y_l.\end{aligned}$$

Since the $\{y_l\}$ are conditionally independent, the conditional variance of S is small and so S can be approximated by its conditional expectation. Thus

$$S = \frac{1}{m} \sum_{l=1}^{m} y_l \approx \frac{1}{m} \sum_{l=1}^{m} \mathrm{E}(y_l \mid T_0 + \tau_l) \approx D + 2\left(\sum_{j=1}^{4} p_j \lambda_j\right)\bar{\tau},$$

where

$$\bar{\tau} = \frac{1}{m} \sum_{l=1}^{m} \tau_l.$$

So to estimate D, we need to estimate

$$\Delta = 2\left(\sum_{j=1}^{4} p_j \lambda_j\right)\bar{\tau}.$$

If we have intrapopulation data, say n_1 sequences from population 1 and n_2 sequences from population 2, then all pairwise comparisons between the two populations should be considered. Thus

$$\bar{S} = \frac{1}{n_1 n_2} \sum_{i=1}^{n_1} \sum_{i'=1}^{n_2} \left(\frac{1}{m} \sum_{l=1}^{m} y_{ii',l}\right) \approx D + \bar{\Delta},$$

where

$$\bar{\Delta} = 2\left(\sum_{j=1}^{4} p_j \lambda_j\right)\frac{1}{n_1 n_2} \sum_{i=1}^{n_1} \sum_{i'=1}^{n_2} \left(\frac{1}{m} \sum_{l=1}^{m} \tau_{ii',l}\right).$$

The quantities $y_{ii',l}$ and $\tau_{ii',l}$ are for the ith sequence from population 1 and the ith sequence from population 2 ($1 \leq i \leq n_1$, $1 \leq i' \leq n_2$). Since the $\{\tau_{ii',l}\}$ are identically distributed,

$$\mathrm{E}(\bar{\Delta}) = 2\left(\sum_{j=1}^{4} p_j \lambda_j\right)\mathrm{E}(\tau),$$

where $\mathrm{E}(\tau)$ is the mean of $\tau_{ii',l}$. The variance of $\bar{\Delta}$ is more difficult to calculate and one must use arguments similar to those used by Weir and Basten.

To estimate $\bar{\Delta}$ we use the intrapopulation data. If S_i is the average number of pairwise differences among the n_i sequences in population i ($i = 1, 2$), then, using the same arguments as above, we can show that $\mathrm{E}(S_i) \approx \mathrm{E}(\bar{\Delta})$. Thus, an unbiased estimate of $\bar{\Delta}$ is

$$\hat{\Delta} = \frac{S_1 + S_2}{2}.$$

The accuracy of the estimate depends on the variances of $\bar{\Delta}$ and $\hat{\Delta}$. Three things influence these variances: the within-population sample sizes, the rate of recombination in the region, and the size of the region. Increasing the sample sizes does not drive the variance of $\bar{\Delta}$ and $\hat{\Delta}$ to zero as one might expect (Hudson, 1983), and so taking larger samples does not continue to improve the estimate. The higher the rate of recombination in the region, the less correlated the m ancestral trees and the better the estimate. The worst case is if the region is completely linked. Increasing the size of the region sequenced improves the estimate because the more distant two nucleotides are, the less correlated are their ancestral trees. Thus small samples of large regions of DNA are more informative than large samples of small regions of DNA. This observation was also made by Weir and Basten.

References

Gojobori, T., Ishii, K., and Nei, M. (1982). Estimation of average number of nucleotide substitutions when the rate of substitution varies with nucleotide. *Journal of Molecular Evolution* **18,** 414–423.

Hudson, R. R. (1983). Properties of a neutral allele model with intragenic recombination. *Theoretical Population Biology* **23,** 183–201.

Jukes, T. H. and Cantor, C. R. (1969). Evolution of protein molecules. In *Mammalian Protein Metabolism*, H. N. Munro (ed.), 21–123. New York: Academic Press.

Kimura, M. (1980). A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. *Journal of Molecular Evolution* **16,** 111–120.
Kimura, M. (1981). Estimation of evolutionary distances between homologous nucleotide sequences. *Proceedings of the National Academy of Science U.S.A.* **78,** 454–458.
Nei, M. (1987). *Molecular Evolutionary Genetics.* New York: Columbia University Press.
Takahata, N. and Kimura, M. (1981). A model of evolutionary base substitutions and its application with special reference to rapid change of pseudogenes. *Genetics* **98,** 641–657.

Simon Tavaré (Departments of Mathematics and Biological Sciences, University of Southern California, Los Angeles, California 90089-1113, U.S.A.);
Ron S. Lundstrom (Department of Mathematics, University of Utah, Salt Lake City, Utah 84112, U.S.A.)

1. Introduction

The authors have made a timely contribution to the statistical literature on the analysis of samples of DNA sequences. Their analysis focuses on the role of ancestry in the generation of data, and as such falls under the umbrella of what might be called "coalescent methods." Our comments are intended to be a brief introduction to this extraordinarily powerful technique that has played and will continue to play an important role in the analysis of neutral population genetics models, both from a theoretical point of view and a more statistical one.

2. Coalescents

We will assume that our populations are large, and that time has been rescaled to units of $2N$ generations, N being the (effective) population size (which may differ in the two populations). Given these assumptions, the genealogy of a sample of n genes from one of the populations is adequately described by the coalescent [Kingman (1982); see Tavaré (1984) for a review]. Under this model, the time during which the sample has j distinct ancestors has an exponential distribution with mean $2/[j(j-1)]$, these times being independent for different j. Think of the ancestry of the sample as going through periods $T_n, T_{n-1}, \ldots$ with $n, n-1, \ldots$ distinct ancestors, eventually tracing back to a single common ancestor (or perhaps to a random number of ancestors if the time-depth in the sample is known).

It is often convenient to think of the coalescent as producing a collection of inverted binary trees, each tree corresponding to a group of individuals who share a common ancestor. The order in which nodes coalesce in these trees is random, reflecting the reproductive symmetry inherent in the model.

The coalescent may be used to describe the ancestry of the sample. The effects of mutation are modelled by superimposing a mutation mechanism on the branches of the coalescent trees. This mutation mechanism can be extremely general, but for the present purpose it is sufficient to focus on the case in which only substitutions of one base for another are allowed. We think of the DNA sequences as comprising s completely linked sites. At each site, substitutions occur at the points of a Poisson process of rate $\theta/2$. (θ is a compound parameter. If u is the probability of a substitution at a given base in a given generation then $\theta = \lim_{N\to\infty} 4Nu$.) It is convenient to label the bases A, C, G, T as 1, 2, 3, and 4 respectively. When a substitution occurs at a base of type i, it is replaced by a base of type j with probability p_{ij}; $p_{ii} > 0$ is allowed here. At a given site, the substitution processes in different parts of the tree are independent of one another. Conditional on the

branch lengths of the coalescent, the substitution processes at different sites are also independent.

Denote by $\mathbf{P}$ the stochastic matrix with elements $\{p_{ij}\}$. The model may be rephrased by saying that the substitution process in a single branch (of, say, length t) is a continuous-time Markov process with generator $\mathbf{Q} = (\theta/2)(\mathbf{P} - \mathbf{I})$, $\mathbf{I}$ being the identity matrix. The probability that a site of type i at the beginning of the branch is of type j at time t is then $p_{ij}(t)$, the ijth element of $e^{\mathbf{Q}t}$. A special case familiar to population geneticists (cf. Griffiths, 1980a; Felsenstein, 1981) is that in which $\mathbf{P}$ is the independent trials process with identical rows $\boldsymbol{\pi} = (\pi_1, \pi_2, \pi_3, \pi_4)$, say. In this case,

$$p_{ij}(t) = e^{-\theta t/2}\delta_{ij} + (1 - e^{-\theta t/2})\pi_j. \tag{2.1}$$

The substitution process here allows "dummy substitutions," the replacement of a given type by itself. The actual number of substitutions in the branch of length t is the number of *changes of state* made by the substitution process, and (for either large t or a stationary initial distribution, $\boldsymbol{\pi}$) the mean of this number is $\kappa = (\theta t/2)\sum_i \pi_i(1 - p_{ii})$. For the model (2.1), this reduces to

$$\kappa = \frac{\theta t}{2}\left(1 - \sum_i \pi_i^2\right). \tag{2.2}$$

Weir and Basten's model may be obtained from (2.1) by setting $\boldsymbol{\pi} = (\frac{1}{4}, \frac{1}{4}, \frac{1}{4}, \frac{1}{4})$, and then using (2.2) to match up the time scales. In what follows, θ should be replaced by $4\theta/3$ to correspond to their scaling.

3. The Divergence of Two Populations

The data comprise samples of DNA sequences from two populations that have been isolated for a time t. (Time is measured in the coalescent time scale. In practice, if the generation length and population sizes of the two populations are sufficiently different, it might be necessary to have different elapsed times for the two populations. We do not consider this complication further here.) Interest focuses on assessing sequence similarity between and within the populations.

It should be clear that in any analysis of this question, an important role is played by the assumptions made about the population at the time of isolation. Consider, for example, a sample of one sequence taken from each population, and focus attention on a given homologous site in those sequences. If we assume, as do Weir and Basten, that the ancestors of the sample are distinct and that the types of those ancestors may be assigned independently with probability x_i of type i, then the probability q_t that the two individuals are identical at that site is

$$q_t = \sum_i \sum_{l,m} x_m x_l p_{li}(t) p_{mi}(t).$$

For the model (2.1), this reduces to

$$q_t = \sum_i (x_i e^{-\theta t/2} + (1 - e^{-\theta t/2})\pi_i)^2, \tag{3.1}$$

a result that gives some indication of the relative importance of the initial frequencies $\mathbf{x}$ and stationary frequencies $\boldsymbol{\pi}$.

On the other hand, the ancestors may themselves be related by descent (cf. Watterson, 1985). If we assume that these ancestors form a stationary sample of size 2 from *their* population then the coalescent method gives the probability $x(l, m)$ of their being type

l and m respectively as

$$x(l, m) = \int_0^\infty e^{-s} \sum_i \pi_i p_{il}(s) p_{im}(s)\, ds. \tag{3.2}$$

This is obtained by conditioning on exponential (mean 1) length(s) of the time back to their common ancestor, and superimposing the effects of substitutions. For model (2.1), this reduces to

$$x(l, m) = \begin{cases} (\theta\pi_l)(\theta\pi_m)/[\theta(1 + \theta)], & l \neq m \\ (\theta\pi_m)(\theta\pi_m + 1)/[\theta(1 + \theta)], & l = m, \end{cases} \tag{3.3}$$

a classical result that can be derived in several other ways. The probability q_t is then given by

$$q_t = \sum_i \sum_{l,m} x(l, m) p_{li}(t) p_{mi}(t),$$

or

$$q_t = \sum_i \pi_i^2 + e^{-\theta t} \frac{(1 - \sum \pi_i^2)}{1 + \theta} \tag{3.4}$$

for the model (2.1).

In contrast, we may calculate the probability Q_t that two individuals chosen at random from one population are identical at a given site. For Weir and Basten's ancestral regime, this may be calculated in terms of the quantities $S_{xx} = \sum x_i^2$, $S_{x\pi} = \sum x_i \pi_i$, $S_{\pi\pi} = \sum \pi_i^2$, as

$$Q_t = \frac{1 + \theta S_{\pi\pi}}{1 + \theta} + e^{-\theta t/2} \frac{2\theta(S_{x\pi} - S_{\pi\pi})}{2 + \theta} + e^{-(\theta+1)t} \times \frac{\theta^2(S_{xx} - 2S_{x\pi} + S_{\pi\pi}) + \theta(3S_{xx} - 2S_{x\pi} - 1) + 2(S_{xx} - 1)}{(1 + \theta)(2 + \theta)}. \tag{3.5}$$

When $\boldsymbol{\pi}$ and $\mathbf{x}$ are identical, this reduces to

$$Q_t = \frac{1 + \theta S_{\pi\pi}}{1 + \theta} + e^{-(\theta+1)t} \frac{(S_{\pi\pi} - 1)}{1 + \theta}. \tag{3.6}$$

On the other hand, for the identity-by-descent regime we obtain

$$Q_t = \frac{1 + \theta S_{\pi\pi}}{1 + \theta}. \tag{3.7}$$

A comparison of (3.1) and (3.5) with (3.4) and (3.7) shows that the two ancestral models are qualitatively quite different. Further analysis of the role of the founding population seems both interesting and important.

4. Lines of Descent

The derivation of analogues of Weir and Basten's quantities T_t and G_t for more general substitution processes is not straightforward. For the independent substitution model (2.1), progress may be made by making use of a line-of-descent process closely related to the coalescent. A sample of individuals taken at time t from the population may be divided into new and old classes (Griffiths, 1980b; Watterson, 1984). A group of individuals is in the same old class if they can be traced back to a common ancestor at time 0, with no intervening mutations. The genetic type of such a group is that of the ancestor. A group of

individuals is in the same new class if they have a most recent common ancestor at or before time 0, that ancestor himself being a mutant. The distribution of the number of lines of descent in a sample of size n is known (cf. Griffiths, 1980b; Tavaré, 1984) explicitly.

The probability $h_{nk}(t)$ that there are k lines of descent is

$$h_{nk}(t) = \sum_{i=k}^{n} \frac{e^{-d_i t}(-1)^{i-k}(2i + \theta - 1)(k + \theta)_{(i-1)} n_{[i]}}{k!(i - k)!(n + \theta)_{(i)}}, \tag{4.1}$$

where $k = 0, 1, \ldots, n$; $a_{[i]} = a(a - 1) \cdots (a - i + 1)$; $a_{(i)} = a(a + 1) \cdots (a + i - 1)$ and $d_i = i(i + \theta - 1)/2$. Watterson (1984) analyses the joint distribution of the number of genes belonging to different old and new classes. We may exploit Watterson's results for our purposes by observing that individuals in a given new class have the same type (that base being type i with probability π_i, $i = 1, 2, 3, 4$) and that different new classes are given a type independently of each other. For example, this allows us to compute the probability T_t that three randomly chosen individuals have identical bases at a given site in the form

$$T_t = \sum_{j=0}^{3} h_{3j}(t)\gamma_j, \tag{4.2}$$

where

$$\begin{aligned}
\gamma_0 &= \sum_i \frac{(\theta\pi_i)(\theta\pi_i + 1)(\theta\pi_i + 2)}{\theta(\theta + 1)(\theta + 2)}, \\
\gamma_1 &= \sum_i \frac{x_i(\theta\pi_i + 1)(\theta\pi_i + 2)}{(\theta + 1)(\theta + 2)}, \\
\gamma_2 &= \sum_i \frac{x_i^2(2 + \theta\pi_i)}{(\theta + 2)}, \\
\gamma_3 &= \sum_i x_i^3.
\end{aligned} \tag{4.3}$$

Similar but more involved combinatorial arguments may be used to derive an analogous formula for G_t, the probability that a random sample of 4 genes has two identical pairs. G_t may be computed from expressions such as

$$\Pr(\text{sample of 4 has 2 } A, \text{ 2 } C \text{ bases}) = \sum_{j=0}^{4} h_{4j}(t)\delta_j,$$

where

$$\begin{aligned}
\delta_0 &= \frac{6(\theta\pi_1)(\theta\pi_2)(\theta\pi_1 + 1)(\theta\pi_2 + 1)}{\theta(\theta + 1)(\theta + 2)(\theta + 3)}, \\
\delta_1 &= \frac{3(1 + \theta\pi_1)(1 + \theta\pi_2)(x_1\theta\pi_2 + x_2\theta\pi_1)}{(\theta + 1)(\theta + 2)(\theta + 3)}, \\
\delta_2 &= \frac{4x_1x_2(1 + \theta\pi_1)(1 + \theta\pi_2) + x_1^2\theta\pi_2(1 + \theta\pi_2) + x_2^2\theta\pi_1(1 + \theta\pi_1)}{(\theta + 2)(\theta + 3)}, \\
\delta_3 &= \frac{3x_1x_2[x_1(1 + \theta\pi_2) + x_2(1 + \theta\pi_1)]}{\theta + 3}, \\
\delta_4 &= 6x_1^2x_2^2,
\end{aligned}$$

and

$$\Pr(\text{sample of 4 has 4 } A \text{ bases}) = \sum_{j=0}^{4} h_{4j}(t)\epsilon_j,$$

where

$$\epsilon_0 = \frac{\theta\pi_1(\theta\pi_1 + 1)(\theta\pi_1 + 2)(\theta\pi_1 + 3)}{\theta(\theta + 1)(\theta + 2)(\theta + 3)},$$

$$\epsilon_1 = \frac{x_1(1 + \theta\pi_1)(2 + \theta\pi_1)(3 + \theta\pi_1)}{(\theta + 1)(\theta + 2)(\theta + 3)},$$

$$\epsilon_2 = \frac{x_1^2(2 + \theta\pi_1)(3 + \theta\pi_1)}{(\theta + 2)(\theta + 3)},$$

$$\epsilon_3 = \frac{x_1^3(3 + \theta\pi_1)}{\theta + 3},$$

$$\epsilon_4 = x_1^4.$$

Related aspects of such sampling probabilities appear in Lundstrom (unpublished Ph.D. thesis, University of Utah, 1990).

5. Discussion

The feature of Weir and Basten's model that makes it tractable is its simplicity; of course, this is precisely the feature that makes it rather unrealistic. As they point out, "there is room for extensions in the direction of greater biological realism." We believe that the coalescent machinery described briefly here provides a very flexible and general method for such extensions. The method, which separates the reproductive process from the mutation mechanism, allows a great variety of mutation schemes to be studied easily. For example, it is possible to allow for different mutation structure in different regions of the sequence, dependence between different sites in the sequence, and invariable regions. The theory we have described here is quite general. However, explicit results for even the simplest generalisations of the Jukes–Cantor model, as typified by the independent trials model described earlier, lead to rather intractable algebraic problems. It seems unrealistic to expect analytical solutions for such complicated problems. However, the coalescent technique can be adapted to generate either recursive systems for probabilities of interest, or simulation results; cf. Lundstrom's unpublished Ph.D. thesis cited previously. Both these areas will be important as we strive for more biological realism. These methods can also be extended to account for recombination (which generates correlated trees for different sites) and for some forms of selection. A very nice review of these aspects appears in Hudson (1990).

ACKNOWLEDGEMENTS

The authors were supported in part by NSF Grant DMS 88-03284 and NIH Grant GM-41746. Simon Tavaré would like to thank Dr R. C. Griffiths for several helpful comments.

REFERENCES

Felsenstein, J. (1981). Evolutionary trees from DNA sequences: A maximum likelihood approach. *Journal of Molecular Evolution* **17**, 368–376.

Griffiths, R. C. (1980a). Allele frequencies in multidimensional Wright–Fisher models with general symmetric mutation structure. *Theoretical Population Biology* **17,** 51–70.

Griffiths, R. C. (1980b). Lines of descent in the diffusion approximation of neutral Wright–Fisher models. *Theoretical Population Biology* **17,** 37–50.

Hudson, R. R. (1990). Gene genealogies and the coalescent process. In *Oxford Surveys in Evolutionary Biology*, Vol. 7, D. Futuyma and J. Antonovics (eds). Oxford: Oxford University Press.

Kingman, J. F. C. (1982). On the genealogy of large populations. *Journal of Applied Probability* **19A,** 27–43.

Tavaré, S. (1984). Line-of-descent and genealogical processes, and their application in population genetics models. *Theoretical Population Biology* **26,** 119–164.

Watterson, G. A. (1984). Lines of descent and the coalescent. *Theoretical Population Biology* **26,** 77–92.

Watterson, G. A. (1985). The genetic divergence of two populations. *Theoretical Population Biology* **27,** 298–317.