Evolutionary Change of Restriction Cleavage Sites and Phylogenetic Inference for Man and Apes[1]

Masatoshi Nei and Fumio Tajima

Center for Demographic and Population Genetics, University of Texas at Houston

A mathematical theory for the evolutionary change of restriction endonuclease cleavage sites is developed, and the probabilities of various types of restriction-site changes are evaluated. A computer simulation is also conducted to study properties of the evolutionary change of restriction sites. These studies indicate that parsimony methods of constructing phylogenetic trees often make erroneous inferences about evolutionary changes of restriction sites unless the number of nucleotide substitutions per site is less than 0.01 for all branches of the tree. This introduces a systematic error in estimating the number of mutational changes for each branch and, consequently, in constructing phylogenetic trees. Therefore, parsimony methods should be used only in cases where nucleotide sequences are closely related. Reexamination of Ferris et al.'s data on restriction-site differences of mitochondrial DNAs does not support Templeton's conclusions regarding the phylogenetic tree for man and apes and the molecular clock hypothesis. Templeton's claim that Nei and Li's method of estimating the number of nucleotide substitutions per site is seriously affected by parallel losses and loss-gains of restriction sites is also unsupported.

Introduction

In the past several years, a number of authors (Avise et al. 1979, 1983; Brown and Simpson 1981; Ferris et al. 1981*a*, 1981*b*, 1983*a*, 1983*b*; Yonekawa et al. 1981; Cann et al. 1982; Nei 1982) have studied phylogenetic relationships of closely related organisms by using restriction endonuclease cleavage-site data for mitochondrial DNA (mtDNA). In these studies, Farris's (1970) parsimony method has often been used, although some authors prefer distance matrix methods (see Sneath and Sokal 1973). Recently, Templeton (1983*a*, 1983*b*) introduced a modified version of the parsimony method, taking into account several properties of the evolutionary change of restriction sites. He claimed that this new method is more powerful in making phylogenetic inferences than are methods based on genetic distances computed by Nei and Li's (1979) method. Applying this method to restriction-site data from man and apes, he concluded that the phylogenetic tree constructed by Ferris et al. (1981*b*) for these organisms is significantly better than several other alternative trees. He also stated that the Nei and Li distance is seriously affected by parallel losses or gain-losses of restriction sites. (Templeton used the word convergent rather than parallel, but corresponding losses of a restriction site in two lineages cannot make two sequences more alike than they were.) Furthermore, he proposed

[1] Key words: Restriction sites, phylogenetic trees, molecular clock, parsimony methods, evolutionary distance.

Address for correspondence and reprints: Dr. Masatoshi Nei, Center for Demographic and Population Genetics, University of Texas at Houston, P.O. Box 20334, Houston, Texas 77225.

Mol. Biol. Evol. 2(3):189–205. 1985.

0737-4038/85/0203-0075$02.00

a method of testing the molecular clock hypothesis with restriction-site data, and, applying this method to data for man and apes, he rejected this hypothesis.

However, the theoretical basis of the parsimony method for restriction-site data is not well established, and it is not clear how reliable the reconstructed tree is. Felsenstein (1978) has studied several cases in which certain parsimony methods fail to give the true tree even if an infinite number of data are used. Although his treatment is deterministic, it warns against an uncritical use of parsimony methods. For restriction-site data, it is possible to conduct a detailed mathematical study of the evolutionary change and to evaluate the possible amount of error introduced into phylogeny construction. In this paper, we present results of our study of this problem and examine several of Templeton's conclusions concerning the evolution of man and apes, the molecular clock hypothesis, and Nei and Li's distance measure.

Mathematical Formulation

Templeton's (1983*a*) algorithm of phylogeny construction consists of three rules. (1) A phylogeny (or evolutionary event) that requires the minimum number of mutational changes is preferred to other possible phylogenies (or evolutionary events). In this case, Farris's (1970) parsimony method is used. (2) When an observed pattern of restriction-site "polymorphism" among the species examined can be explained by the same number of mutational changes in two different ways, the one with parallel losses or gain-losses is preferred to that with parallel gains or loss-gains. This rule is based on Templeton's (1983*b*) study in which parallel losses and gain-losses were shown to occur more frequently than parallel gains and loss-gains. (Polymorphism is placed in quotes to indicate that the sequence differences are between species rather than within species.) (3) The consistency of a phylogenetic network inferred from a restriction enzyme with each of several alternative phylogenetic trees is examined, and the tree that is consistent with the networks for the largest number of enzymes is chosen as the probable tree (Estabrook et al.'s [1976] procedure).

As an example, consider the hypothetical phylogenetic tree in figure 1, in which species A and B have a restriction site at a given DNA site, whereas species C, D, and E do not. This pattern of restriction-site "polymorphism" among the five species can be explained either by two independent gains of restriction sites (a; fig. 1) or by three independent losses (b; fig. 1). According to Farris's parsimony method, we must choose a, because a has fewer mutational events than b. In reality, however, the probability of occurrence of b is not always lower than that of a, as will be shown below. That is, under certain conditions, the parsimony method has a high probability of making an erroneous inference about the evolutionary change of restriction sites. To see the conditions under which this occurs, however, we must first establish some rules regarding the evolutionary change of restriction sites.

Basic Theory

Consider a restriction endonuclease with a particular recognition sequence of r nucleotides and denote by W_i a sequence of r nucleotides that differs from the recognition sequence by i nucleotides. In particular, W_0 denotes a DNA sequence that is identical with the recognition sequence, i.e., a restriction site. For simplicity, we assume that the rate of nucleotide substitution is the same for all nucleotides and that the frequencies of the four nucleotides A, T, G, and C are in equilibrium

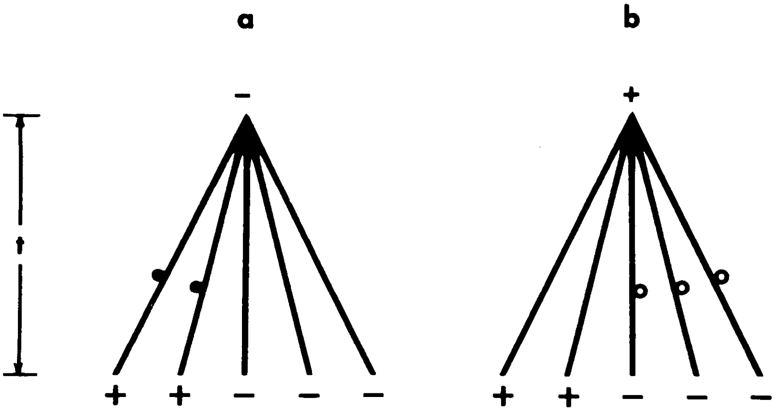

FIG. 1.—Two possible explanations of the observed pattern of restriction-site "polymorphism" among the five species A, B, C, D, and E. A plus sign denotes the presence of a given restriction site; a minus sign denotes the absence of the restriction site. Darkened circles indicate gains of a given restriction site; unshaded circles represent losses of the restriction site. t = evolutionary time.

and equal to ¼. (See Discussion for the effect of violation of this assumption.) Under this assumption, the probability that a randomly chosen sequence of r nucleotides is W_i is given by

$$a_i = \binom{r}{i}(3/4)^i(1/4)^{r-i} = \binom{r}{i}3^i/4^r. \qquad (1)$$

We denote by λ the rate of nucleotide substitution per site per year. The probability that a nucleotide at a given site at time t (measured in years) is the same as that at t = 0 is then

$$p(t) = 1/4 + 3/4\, e^{-(4/3)\lambda t} \qquad \text{(Nei and Li 1979).} \qquad (2)$$

The probability that a nucleotide at a given site at time t is different from the nucleotide at time t_0 is

$$\begin{aligned} q(t) &= [1 - p(t)]/3 \\ &= 1/4 - 1/4\, e^{-4/3\lambda t}. \qquad (3) \end{aligned}$$

Let us now consider the change of W_i to W_j. This happens only when k of the i nucleotides that are different from those of the recognition sequence change to the latter nucleotides and exactly r − j − k of the r − i nucleotides that are identical with the recognition nucleotides remain unchanged, the remaining r − i − (r − j − k) = j + k − i nucleotides changing to different nucleotides. Therefore, the probability that W_i becomes W_j during t years is

$$v_{ij}(t) = \sum_k \binom{i}{k} q(t)^k[1 - q(t)]^{i-k} \times \binom{r-i}{r-j-k} p(t)^{r-j-k}[1 - p(t)]^{j+k-i}, \qquad (4)$$

where $k \le \min(i, r - j)$ and, if $i \ge j$, then $k \ge i - j$ or, if $i < j$, then $k \ge 0$. Here, min (a, b) indicates the smaller value of a and b. When $i = 0$, equation (4) reduces to

$$v_{0j}(t) = \binom{r}{r-j}[1 - p(t)]^j p(t)^{r-j}. \tag{5}$$

When $j = 0$, it becomes

$$v_{i0}(t) = q(t)^i p(t)^{r-i}. \tag{6}$$

From equation (4), we have the following equation,

$$a_i v_{ij}(t) = a_j v_{ji}(t). \tag{7}$$

Thus, the expected number of W_i sequences changing to W_j's during t years is equal to that of W_j's changing to W_i's.

Evolutionary Changes of Restriction Sites

Let us now consider the probabilities of 10 basic types of restriction-site changes, which are shown in figure 2. The probability of occurrence (or the expected number of occurrences per site of r nucleotides) of evolutionary relationship a (fig. 2) is

$$\begin{aligned} P_a &= a_0 \sum_{j=1}^{r} v_{0j}(t) = a_0[1 - v_{00}(t)] \\ &= a_0[1 - p(t)^r], \end{aligned} \tag{8}$$

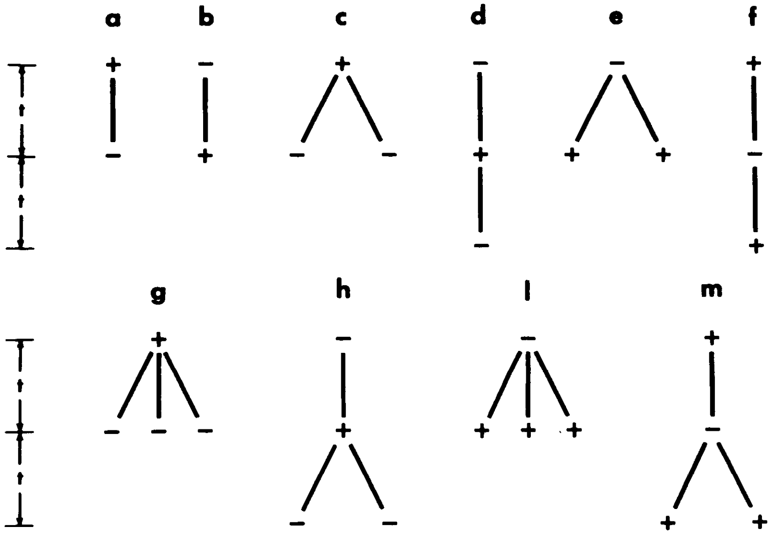

FIG. 2.—Ten different types of restriction-site changes. Symbols are those used in fig. 1.

where $a_0 = (1/4)^r$. The probability of occurrence of relationship b (fig. 2) is

$$P_b = \sum_{i=1}^{r} a_i v_{i0}(t) = \sum_{i=1}^{r} a_0 v_{0i}(t) = a_0[1 - p(t)^r]. \tag{9}$$

Therefore, we have

$$P_b = P_a. \tag{10}$$

This indicates that the probability of loss of extant restriction sites is equal to the probability of gain of new sites. This is obviously so, since at equilibrium the expected number of restriction sites must remain the same. It should, however, be noted that if we consider a particular site of r nucleotides in a DNA sequence, the probability of loss of an extant restriction site is much higher than that of gain of a site. When $r\lambda t \ll 1$, the probability of loss is approximately $r\lambda t$, whereas the probability of gain is approximately $r\lambda t(1/4)^r$.

The probabilities of the other relationships in figure 2 can be obtained in the same way. They become

$$P_c = P_d = a_0[\sum_{i=1}^{r} v_{0i}(t)]^2 = a_0[1 - p(t)^r]^2, \tag{11}$$

$$P_e = P_f = a_0 \sum_{i=1}^{r} v_{0i}(t)v_{i0}(t) = a_0[p(2t)^r - p(t)^{2r}], \tag{12}$$

$$P_g = P_h = a_0[\sum_{i=1}^{r} v_{0i}(t)]^3 = a_0[1 - p(t)^r]^3, \tag{13}$$

$$P_l = P_m = a_0 \sum_{i=1}^{r} v_{0i}(t)v_{i0}(t)^2 = a_0(\{[1 - p(t)]q(t)^2 + p(t)^3\}^r - p(t)^{3r}), \tag{14}$$

where P_i denotes the probability of obtaining relationship i. Equation (11) indicates that the probability of parallel losses is equal to that of gain-losses, whereas equation (12) shows that the probability of parallel gains is equal to that of loss-gains.

Numerical values of the above probabilities for various λt values are given in table 1. It is clear that the probability of parallel gains (P_e) or loss-gain (P_f) is always much smaller than that of parallel losses (P_c) or gain-losses (P_d). This is in agreement with Templeton's conclusion. Why is tree d more probable than tree f? The reason is as follows: In the case of d, there is no restriction site initially, and a new restriction site can be created at any potential position in the DNA sequence during the first t years. The probability of occurrence of this event is equal to the probability of subsequent loss of the restriction site during the next t years. Therefore, this probability is given by equation (11). In the case of f, however, a restriction site is lost during the first t years, but during the second t years the restriction site is restored at exactly the same position held by the previous one. The probability of occurrence of the latter event is generally much lower than the probability of loss of a restriction site. Therefore, P_f is much lower than P_d. Similarly, P_e is much lower than P_c because, in tree e, a restriction site must be created at the same DNA position for the two lineages considered.

Table 1
Probabilities of Having 12 Different Types of Restriction-Site Changes Shown in Figure 2 (a, b, —, and m) and Figure 1 (a and b)

	λt				
Case	0.001	0.005	0.01	0.05	0.1
$P_a = P_b$	1.5×10^{-6}	7.2×10^{-6}	1.4×10^{-5}	6.3×10^{-5}	1.1×10^{-4}
$P_c = P_d$	8.7×10^{-9}	2.1×10^{-7}	8.3×10^{-7}	1.6×10^{-5}	4.9×10^{-5}
$P_e = P_f$	4.8×10^{-10}	1.2×10^{-8}	4.4×10^{-8}	7.0×10^{-7}	1.6×10^{-6}
$P_g = P_h$	5.2×10^{-11}	6.3×10^{-9}	4.8×10^{-8}	4.2×10^{-6}	2.2×10^{-5}
$P_l = P_m$	1.6×10^{-13}	2.0×10^{-11}	1.4×10^{-10}	8.8×10^{-9}	3.1×10^{-8}
$P_{(a)}$	4.8×10^{-10}	1.1×10^{-8}	4.3×10^{-8}	6.7×10^{-7}	1.5×10^{-6}
$P_{(b)}$	5.2×10^{-11}	5.9×10^{-9}	4.3×10^{-8}	2.3×10^{-6}	6.6×10^{-6}

NOTE.—P_a is the probability of type a change, etc. A recognition sequence of six nucleotides is assumed.

Table 1 shows that the probability of three parallel losses (P_g) or of one-gain/two-losses (P_h) is not necessarily smaller than that of two parallel gains (P_e). Indeed, when $\lambda t \geq 0.01$, P_g/P_h is greater than P_e. This indicates that the parsimony method introduces a systematic error in phylogeny construction unless λt is very small. Templeton (1983*b*) recognized this problem but concluded that if $\lambda t < 0.05$, the error introduced in a parsimony tree is small. However, his conclusion is derived from a study of the case of one or two evolutionary lineages; if more than two lineages are considered, a lower criterion ($\lambda t < 0.01$) is necessary.

The probabilities of having evolutionary events a and b in figure 1 can be obtained in the same way. They become

$$P_{(a)} = a_0 \sum_{i=1}^{r} v_{0i}(t)v_{i0}(t)[\sum_{j=1}^{r} v_{ij}(t)]^3,$$

$$P_{(b)} = a_0 v_{00}(t)^2[\sum_{i=1}^{r} v_{0i}(t)]^3.$$

The numerical values of these probabilities are also given in table 1. When λt is small, $P_{(a)}$ is higher than $P_{(b)}$, but when $\lambda t > 0.01$, $P_{(a)}$ is lower than $P_{(b)}$.

Figure 3 shows another type of evolutionary tree, in which the time (t_1) between the first and second splittings of species is not equal to the time (t_2) between the second splitting of species and the present. The restriction-site differences among the three species involved can be explained either by two independent losses of the restriction site (or one gain-loss) (a; fig. 3) or by one gain (b; fig. 3). According to the parsimony principle, we must choose b. The probabilities of obtaining these two events can be computed by using the theory developed above. They become

$$P_a = a_0 v_{00}(t_2) \sum_{i=1}^{r} v_{0i}(t_2) \sum_{j=1}^{r} v_{0j}(2t_1 + t_2) = a_0 p(t_2)^r[1 - p(t_2)^r][1 - p(2t_1 + t_2)^r], \quad (15)$$

$$P_b = a_0 \sum_{i=1}^{r} v_{0i}(t_2) \sum_{j=1}^{r} v_{ij}(t_2) \sum_{k=1}^{r} v_{ik}(2t_1 + t_2) = a_0(1 - v_1 - v_2 + v_3) - P_a, \quad (16)$$

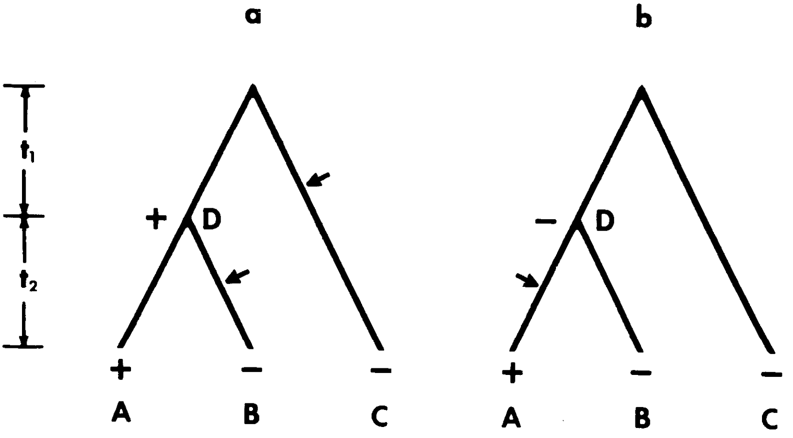

FIG. 3.—Two possible schemes of mutational changes that produce the observed pattern of restriction-site "polymorphism" among the three species A, B, and C. Arrows indicate occurrence of loss or gain of a restriction site. t_1 and t_2 = evolutionary time. In the mutational change (a) there are two possible cases: (1) the ancestral sequence had the restriction site, and this site was lost in species B and C independently (parallel losses); and (2) the ancestral sequence did not have the restriction site, but it appeared in the A-B line during the first t_1 years and later disappeared from species B. These two cases are pooled in equation (15). Similarly, the two possible cases for (b) are pooled in equation (16).

where

$$v_1 = \{[1 - p(t_2)]q(t_2) + p(t_2)^2\}^r,$$

$$v_2 = \{[1 - p(t_2)]q(2t_1 + t_2) + p(t_2)p(2t_1 + t_2)\}^r,$$

$$v_3 = \{[1 - p(t_2)]q(t_2)q(2t_1 + t_2) + p(t_2)^2p(2t_1 + t_2)\}^r.$$

Some numerical values of these probabilities are given in table 2. This table also includes the relative probability of evolutionary event a, i.e., $P'_a = P_a/(P_a + P_b)$. When λt_1 and λt_2 are both small, event a occurs with a much lower probability than event b. However, as λt_1 and λt_2 increase, P'_a increases gradually. Particularly when $\lambda t_1 = 0.1$ and $\lambda t_2 = 0.005$, P'_a is as high as 0.4. That is, event a occurs almost as frequently as event b. It is also noted that even if $\lambda t_1 = 0.05$ and $\lambda t_2 = 0.005$, P'_a is 0.314. Therefore, Templeton's criterion of $\lambda t \leq 0.05$ for making reliable inferences regarding the evolutionary change of restriction sites does not hold.

Computer Simulation

In the above study, we considered only simple cases to examine the adequacy of the parsimony method for restriction-site data. However, the actual amount of error introduced depends on the topology and branch lengths of the tree concerned. We have therefore studied this problem further by using computer simulation. In this simulation, we used five "species" in which the evolutionary tree is similar to that of the five primate species (man, chimpanzee, gorilla, orangutan, and gibbon) studied by Templeton. We assumed that the (unknown) true tree for these species is given by the diagram in figure 4. The expected numbers of nucleotide substitutions per site (λt) used for the branches of this tree are approximately equal to the values observed by Brown et al. (1982). In this simulation, we used "six-base enzymes"

Table 2
Probabilities of Evolutionary Changes of Restriction Sites Shown in Figure 3

λt_1	λt_2	P_a	P_b	P'_a
0.005	0.005	6.1×10^{-7}	7.2×10^{-6}	0.077
	0.01	1.5×10^{-6}	1.4×10^{-5}	0.097
	0.05	1.4×10^{-5}	6.1×10^{-5}	0.186
	0.1	2.9×10^{-5}	1.1×10^{-4}	0.214
0.01	0.005	9.7×10^{-7}	7.1×10^{-6}	0.120
	0.01	2.2×10^{-6}	1.4×10^{-5}	0.136
	0.05	1.6×10^{-5}	6.1×10^{-5}	0.206
	0.1	3.1×10^{-5}	1.1×10^{-4}	0.224
0.05	0.005	3.2×10^{-6}	7.1×10^{-6}	0.314
	0.01	6.4×10^{-6}	1.4×10^{-5}	0.315
	0.05	2.7×10^{-5}	6.1×10^{-5}	0.309
	0.1	4.1×10^{-5}	1.1×10^{-4}	0.282
0.1	0.005	4.9×10^{-6}	7.0×10^{-6}	0.408
	0.01	9.4×10^{-6}	1.4×10^{-5}	0.404
	0.05	3.6×10^{-5}	6.1×10^{-5}	0.368
	0.1	4.9×10^{-5}	1.1×10^{-4}	0.319

NOTE.—A recognition sequence of six nucleotides is assumed.

and assumed that the rate of nucleotide substitution per year (λ) is constant and the same for all nucleotide pairs. We first generated a random sequence of six nucleotides and regarded this as the ancestral sequence (I). We then followed the evolutionary change of this sequence according to the tree in figure 4. For a given

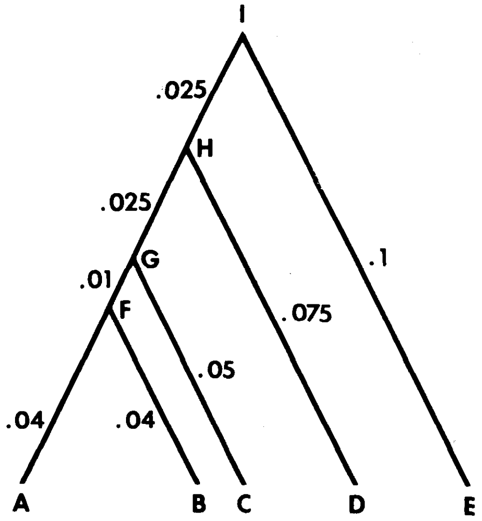

FIG. 4.—Phylogenetic tree according to which computer simulation was done. A, B, C, D, and E are extant species, whereas F, G, H, and I represent ancestral species. The number given for each branch is the expected number of nucleotide substitutions (λt). This phylogenetic tree is close to that for the chimpanzee (A), gorilla (B), man (C), orangutan (D), and gibbon (E) obtained by Ferris et al. (1981*b*) and Brown et al. (1982).

evolutionary time t, each nucleotide remained unchanged with probability p(t) in equation (2) and changed to one of the three remaining nucleotides with probability 1 − p(t). In the latter case, we assumed that the probability of a nucleotide changing to each of the three alternative nucleotides is ⅓. The nucleotide sequences for the five extant species (A, B, C, D, and E) and for the four ancestral species (F, G, H, and I) were recorded. We then regarded the sequence for species A as the recognition sequence for a restriction enzyme and determined whether the other four extant species had the same sequence. (Since nucleotides A, T, G, and C were equally frequent, any sequence of r nucleotides could be defined as a recognition sequence.) When all five extant species had the same recognition sequence (restriction site), those data were discarded as uninformative for constructing a parsimony tree. When some other species had nonrecognition sequences, the evolutionary changes of restriction sites for all branches were examined (see fig. 5 for examples). This produced one replicate of the evolutionary change of restriction sites. The second replicate was obtained from the same replication of computer simulation by changing the recognition sequence from the sequence for species A to that for species B, C, D, or E when the latter was different from that of A. We continued this process until all different nucleotide sequences in the extant species were used as recognition sequences. Since the expected nucleotide frequencies are all ¼ and we are interested only in the nucleotide differences among related species, this procedure is justified. This procedure saves much computer time yet gives essentially the same result as that of Li (1981*a*).

We repeated this procedure 1,000 times and produced 2,601 replicates of the evolutionary change of restriction sites. We then classified the types of restriction-site changes and tabulated their frequencies. There were 102 different types of restriction-site changes observed, and some of the nonparsimonious restriction-site changes were quite frequent. In the example shown in figure 5, species A and B have a restriction site that the other species lack. According to the parsimony principle, this pattern of restriction-site polymorphism among the five species is considered to have occurred by a single restriction-site gain, as shown in a (fig. 5). In practice, however, we observed five different types of changes that produced the same polymorphic pattern. Type a is certainly the most frequent one but does not account for the majority of changes, the proportion of a among all five types being 0.41. Type b (one gain and one loss; fig. 5) and type e (three losses; fig. 5) also

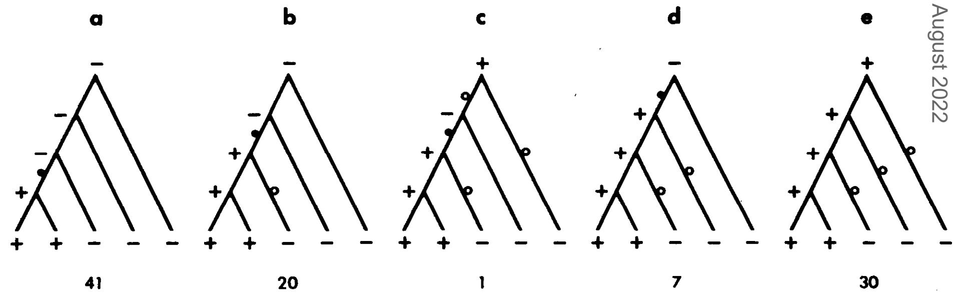

FIG. 5.—Five different types of evolutionary changes of restriction sites that produced the same pattern of restriction-site "polymorphism" among the five species examined in computer simulation. The number below each diagram gives the number of occurrences of the evolutionary change of restriction sites indicated. Symbols are those used in fig. 1.

occur quite frequently. The proportion of type a is expected to decline as the expected length (λt) of branch G-F decreases. Indeed, in another, similar computer simulation in which the λt value for branch G-F was set at 0.005, the proportion of type a (0.24) was smaller than that of type b (0.33) and type e (0.31). (Space limitation prevents discussion of the details of this simulation.) These results confirm our conclusion from the theoretical study that parsimony introduces a systematic error in the inference of evolutionary change of restriction sites. This is true even if rule (ii) of Templeton's algorithm is used.

Because of this erroneous inference, the parsimony method is expected to give an incorrect estimate of the number of mutational changes for some branches of the tree. This is indeed the case, as shown in figure 6. The number given for each branch in a (fig. 6) is the observed number of restriction-site changes in computer simulation, whereas the value in b (fig. 6) is the number estimated by Templeton's maximum parsimony/minimum parallel gains and/or loss-gain method. In most branches, the estimated number is certainly smaller than the observed number, but in branches G-F and H-G it is larger than the observed number. The overestimate of the number of mutational changes for branch G-F occurred because in the parsimony method one mutational change was assigned to branch G-F whenever parallel gains or losses occurred in branches F-A and F-B. The overestimate for branch G-H occurred for a similar reason. Obviously, the probability of occurrence of these events is higher when the λt's for branches F-G and G-H are small compared to those for the other branches. In the other computer simulation mentioned above, the estimated number (92) for branch G-F was 2.4 times larger than the number actually observed (38). This type of error causes an unexpectedly high degree of underestimation for branches F-A, F-B, and G-C. Thus, even the short branches F-A and F-B of the tree in figure 6 show an average underestimation of 13%. The extent of underestimation for branch G-C is even higher, being 26%.

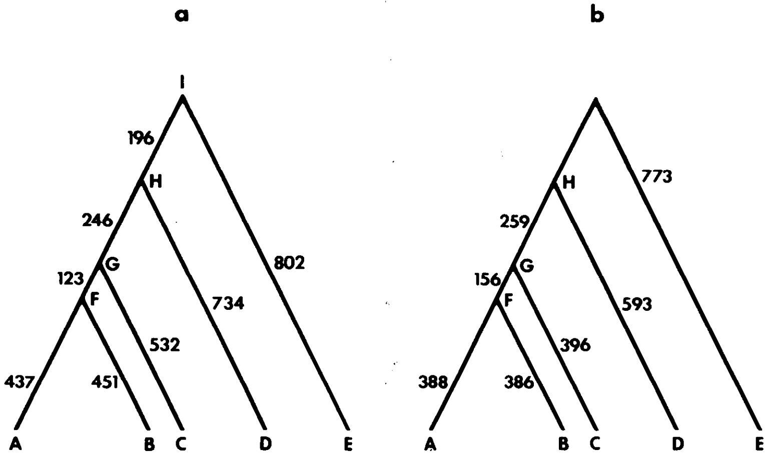

FIG. 6.—(a) Observed numbers of mutational changes for the branches examined in computer simulation and (b) numbers of mutational changes estimated by Templeton's method for the branches examined. In Templeton's method, the numbers of changes for branches I-E and I-H cannot be distinguished, so that an estimate of the sum of the changes for the two branches is given. The observed number for each branch is not the total number of mutational changes that really occurred but the number observed by comparing the two nucleotide sequences at both ends of the branch.

This latter figure is larger than the extent of underestimation for branches H-D (19%) and H-E (23%).

Discussion

Assumption of Equal Substitution Rates

In the present study, we have assumed that the rates of nucleotide substitution among the four nucleotides are equal. In practice, this assumption does not hold for mtDNA (Brown et al. 1982; Aquadro and Greenberg 1983), and the nucleotide frequencies usually deviate from ¼ (Brown 1983). Unequal rates of nucleotide substitution complicate the mathematical formulation of our problem, but the general feature of our findings is expected to remain the same. This is because the expected numbers of gains and losses of restriction sites in a DNA sequence are equal to each other at equilibrium for any type of nucleotide substitution (Nei and Tajima 1983). Indeed, Wen-Hsiung Li (personal communication) has shown that, even when transitional substitutions are much more frequent than transversional substitutions, our conclusion about the evolutionary change of restriction sites is not altered.

Phylogenetic Inference

The present study has shown that the parsimony method can give an erroneous inference about the evolutionary change of restriction sites and that the estimate of the number of mutational changes for a given branch could be larger than the true number. These findings suggest that a systematic error will be introduced in a phylogenetic tree reconstructed by parsimony methods. In our formulation, we have assumed that the rate of nucleotide substitution (λ) is the same for all evolutionary lineages. However, if λ varies with evolutionary lineage, parsimony methods are expected to give even more erroneous trees (Felsenstein 1978).

Let us now examine Templeton's (1983*a*) analysis of primate data in light of our findings. He analyzed Ferris et al.'s (1981*b*) data on restriction sites for mitochondrial DNAs from man, chimpanzee, gorilla, orangutan, and gibbon. He considered six alternative phylogenetic trees and computed the minimum number of mutational changes required to explain the observed pattern of restriction-site polymorphism for each of the alternative trees. He then ranked each hypothetical phylogeny for the 13 informative restriction enzymes used. Using Wilcoxon's signed ranks test, he concluded that his phylogeny 1 (phylogeny A in fig. 7 of the present paper) is significantly better than his phylogenies 2, 3, and 4 (phylogeny B in fig. 7). Particularly in the comparison of phylogenies A and B, the former was always equal to or better than the latter in ranking.

Actually, the superiority of phylogeny A over B in figure 7 is embedded in the parsimony method used, and it can be shown that phylogeny A is never inferior to phylogeny B for any pattern of restriction-site polymorphism. Consider the four different patterns of restriction-site polymorphism among the five species given in figure 7. In polymorphism pattern 1 the chimpanzee and gorilla share a restriction site that the others do not. This pattern can be explained by one mutational change in phylogeny A, but two changes are required in phylogeny B. Therefore, A is better than B. The same conclusion can be obtained for polymorphism pattern 3. In polymorphism pattern 2, neither A nor B is better than the other, because two mutational losses are required in both cases. In polymorphism pattern 4, two mutational changes are required in both trees, but since parallel gains are less

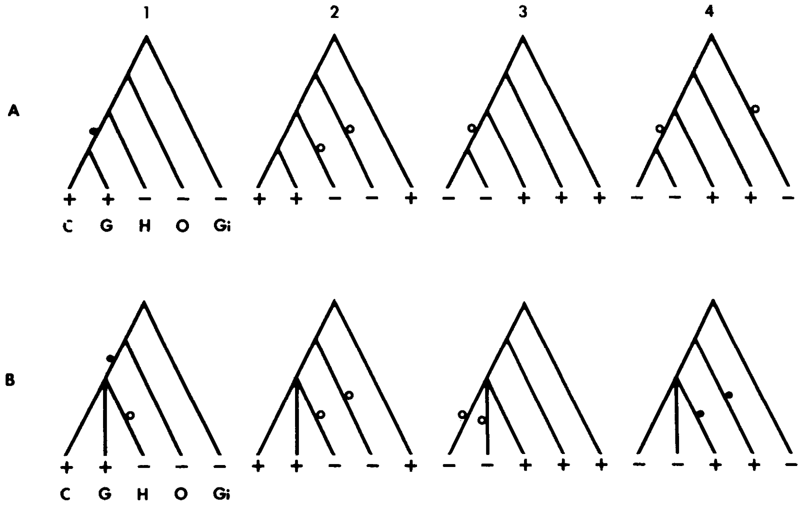

FIG. 7.—Comparisons of four different types of restriction-site polymorphisms (1–4) between phylogenies A and B (Templeton's phylogenies 1 and 4, respectively). C, G, H, O, and Gi stand for the chimpanzee, gorilla, human, orangutan, and gibbon mtDNA sequences, respectively. See text for details. Symbols are those used in fig. 1.

probable than parallel losses, A is given a better score than B according to Templeton's method. Thus, in all four cases A is better than or equal to B in ranking. It can be shown that, in all other polymorphism patterns, the same number of mutational changes is required for both phylogenies A and B. Therefore, phylogeny B cannot be better than A. Templeton's analysis also indicates that for all 13 enzymes examined, A is always better than or equal to B, as expected from this study. Needless to say, this superiority of phylogeny A over phylogeny B does not mean that B cannot be the true phylogeny. Rather, B can be the true one, since the evolutionary change of restriction sites does not necessarily occur in a parsimonious way (see fig. 5).

According to Templeton's (1983*a*) table 2, his tree 1 is also significantly better than his tree 2, in which the chimpanzee is more closely related to man than to the gorilla. In this case, however, he used two enzymes of which the recognition sequences overlap with those of some other enzymes. Namely, the recognition sequences [GT$(^{\mathrm{T}}_{\mathrm{C}})(^{\mathrm{A}}_{\mathrm{G}})$AC] of *Hinc*II include the recognition sequences of *Hpa*I [GTTAAC] and *Sal*I [GTCGAC], whereas those of *Ava*I [C$(^{\mathrm{T}}_{\mathrm{C}})CG(^{\mathrm{A}}_{\mathrm{G}})$G] include the recognition sequences of *Xho*I [CTCGAG] and *Sma*I [CCCGGG]. Therefore, the data from these enzymes are not independent. A simple way to avoid this problem is to eliminate data from *Hinc*II and *Ava*I. If we do this, the difference between trees 1 and 2 is no longer significant. Templeton's table 2 also shows that his tree 1 is significantly better than his tree 3, in which the gorilla is more closely related to man than to the chimpanzee. In this case, his conclusion may be correct, but the possibility that this is also caused by erroneous inferences about the evolutionary changes of restriction sites cannot be ruled out.

In this connection, it should be noted that recent DNA hybridization data (Sibley and Ahlquist 1984) support Templeton's tree 2 rather than his tree 1. Templeton's tree 2 is also supported by protein sequence data (Goodman et al.

1982), chromosome banding patterns (Yunis and Prakash 1982), and a reanalysis of Brown et al.'s (1982) data on mtDNA sequences (Nei et al. 1985).

In Templeton's algorithm, the compatibility principle plays an important role, as mentioned earlier. However, the validity of this principle is also questionable. This is obvious from the comparison of trees A and B in figure 7 for restriction-site polymorphism pattern 1. According to the parsimony principle, this pattern is explained by a single gain of a restriction site. The network inferred from this type of polymorphism is compatible with phylogeny A but not with phylogeny B. Yet, the latter tree, rather than the former, may be the correct one, as noted above.

Despite these comments, it should be noted that parsimony methods are useful for constructing a phylogenetic tree if the λt value for each branch is very small, say, $\lambda t < 0.01$. In this case, some errors can still occur, but the probability of occurrence of errors seems to be small (table 2). Parsimony methods should be particularly useful when one is interested in finding the evolutionary pathways of DNA sequences sampled from the same species. Indeed, Avise et al. (1979, 1983) and Ferris et al. (1983*a,* 1983*b*) used parsimony methods to clarify the evolutionary relationships of polymorphic mtDNAs in *Peromyscus, Geomys,* and *Mus.* In general, parsimony methods seem to be superior to distance matrix methods when λt is small, but as λt increases, distance matrix methods gradually become superior to parsimony methods, as long as the rate of nucleotide substitution remains more or less the same for all evolutionary lineages (Peacock and Boulter 1975; Blanken et al. 1982; Tateno et al. 1982; Nei et al. 1983). Note that there are several distance matrix methods in which varying rates of nucleotide substitution can be taken into account (Fitch and Margoliash 1967; Farris 1977; Klotz and Blanken 1981; Li 1981*b*).

Molecular Clock Hypothesis

For a parsimonious tree, the estimate of the number of mutational changes for a branch is supposed to be equal to or smaller than the true value (Goodman et al. 1974). Because the extent of underestimation of the number of mutational changes is expected to increase as branch length increases, it is difficult to use parsimonious tree estimates for testing the molecular clock hypothesis or the linear relationship between evolutionary time and mutational changes. To avoid this problem, some authors (e.g., Goodman et al. 1974; Langley and Fitch 1974) devised statistical methods of correcting for undetected mutational changes for each branch. However, there has been a great deal of controversy about the adequacy of these methods (e.g., Czelusniak et al. 1978; Holmquist 1978; Tateno and Nei 1978; Nei and Tateno 1979; Kimura 1981*a,* 1981*b*).

Templeton's (1983*a*) test of the molecular clock hypothesis did not include any correction for undetectable changes. He first compared the number of mutational changes in the human line with those in the gorilla and chimpanzee lines, using information for all five species studied. Wilcoxon's signed ranks test showed that the human line had a significantly smaller number of mutational changes. But since there are two species in the pongid line (chimpanzee and gorilla), the human line is expected to have a smaller number of mutational changes in a parsimonious tree. (Templeton was aware of this problem.) Furthermore, the phylogeny used is expected to give an overestimate of mutational changes for a branch between the common ancestor for the human, chimpanzee, and gorilla and that for the chimpanzee and gorilla (see fig. 6). Therefore, his first test is not warranted.

Templeton's second test is better than his first, since the above two problems are avoided by eliminating one of the two pongid species from the analysis. In this test, the human line no longer showed a significantly smaller number of mutational changes than the chimpanzee line. Templeton's table 8 shows that the human line still has a significantly smaller number of changes than the gorilla line. However, this is apparently caused by a computational error. In this table, the value of d_{G-H} for enzyme i is listed as 1, but the data in his table 1 indicate that this should be −1. (If we eliminate the chimpanzee from the analysis, the polymorphism pattern for restriction site 1 of this enzyme can be explained by two gains, two losses, or one gain-loss. According to Templeton's algorithm, we must choose either the second or the third event. We will then have to assume one extra mutation for the human line compared with the gorilla line.) This change in sign makes the rank for this enzyme change from 6 to −6 and leads to the conclusion that the difference in mutational changes between the human and gorilla lines is not statistically significant. Hence, contrary to Templeton's conclusion, the molecular clock hypothesis cannot be rejected from his data. This conclusion is reinforced if we eliminate data for *Hinc*II and *Ava*I, the recognition sequences of which overlap with those of some other enzymes.

Nei and Li's Distance Measure

As mentioned earlier, Templeton (1983*a*, 1983*b*) argued that Nei and Li's (1979) genetic distance measure is seriously affected by parallel losses and gain-losses. He stated that "for $\lambda t = 0.03$ about as many mutations are ignored as are scored in the Nei and Li distance" (1983*b*, p. 167). This criticism is apparently based on his misunderstanding of Nei and Li's theory, since Nei and Li have never ignored parallel losses and gain-losses. What they did ignore are parallel gains, which occur with a small probability. The fact that the effect of parallel gains on the estimate of nucleotide substitutions per site (δ) is negligibly small was later confirmed by Kaplan and Risko (1981), Li (1981*a*), and Nei and Tajima (1983). Templeton's (1983*b*) own computation (his table 2) also supports Nei and Li's assumption. We have now developed a simple method for estimating δ without making this assumption (Nei and Tajima 1983), but this method gives essentially the same value as that obtained by Nei and Li's formula, unless $2\lambda t > 0.3$.

Citing Adams and Rothman's (1982) paper, Templeton also claimed that the value of λt varies with the restriction enzyme used and thus restriction-site data from different enzymes should not be pooled as in the case of the Nei and Li distance. Actually, what Adams and Rothman showed is not that λt varies with the restriction enzyme used but that the distribution and the observed number of restriction sites in DNA sequences are significantly different from those expected under the assumption of equal nucleotide frequencies. It is well known that the nucleotide frequencies are unequal in most mtDNAs and that there is a deficiency of the dinucleotide CG. If we take into account these two factors, the discrepancy between the observed and expected numbers of restriction sites is substantially reduced (F.T., unpublished observation). Yet, the agreement is not always satisfactory. This remaining discrepancy seems to be the result of unequal rates of substitution among the four types of nucleotides. However, the effect of unequal rates of substitution on δ is known to be small unless $\delta > 0.4$ (Nei and Tajima 1983). When $\delta > 0.4$, the restriction-site method of estimating δ is not reliable for any case, because of the large standard error (Li 1981*a*).

Acknowledgements

We thank Wen-Hsiung Li, Clay Stephens, Ellen Prager, and Allan Wilson for their comments on an early draft of this paper. This work was supported by research grants from the National Institutes of Health and the National Science Foundation.

LITERATURE CITED

Adams, J., and E. D. Rothman. 1982. Estimation of phylogenetic relationships from DNA restriction patterns and selection of endonuclease cleavage sites. Proc. Natl. Acad. Sci. USA **79**:3560–3564.

Aquadro, C. F., and B. D. Greenberg. 1983. Human mitochondrial DNA variation and evolution: analysis of nucleotide sequences from seven individuals. Genetics **103**:287–312.

Avise, J. C., R. A. Lansman, and R. O. Shade. 1979. The use of restriction endonucleases to measure mitochondrial DNA sequence relatedness in natural populations. I. Population structure and evolution in the genus *Peromyscus*. Genetics **92**:279–295.

Avise, J. C., J. F. Shapira, S. W. Daniel, C. F. Aquadro, and R. A. Lansman. 1983. Mitochondrial DNA differentiation during the speciation process in *Peromyscus*. Mol. Biol. Evol. **1**:38–56.

Blanken, R. L., L. C. Klotz, and A. G. Himmebusch. 1982. Computer comparison of new and existing criteria for constructing evolutionary trees from sequence data. J. Mol. Evol. **19**:9–19.

Brown, W. M. 1983. Evolution of animal mitochondrial DNA. Pp. 62–88 *in* M. Nei and R. K. Koehn, eds. Evolution of genes and proteins. Sinauer, Sunderland, Mass.

Brown, W. M., E. M. Prager, A. Wang, and A. C. Wilson. 1982. Mitochondrial DNA sequences of primates: tempo and mode of evolution. J. Mol. Evol. **18**:225–239.

Brown, G. G., and M. V. Simpson. 1981. Intra- and interspecific variation of the mitochondrial genome in *Rattus norvegicus* and *Rattus rattus:* restriction enzyme analysis of variant mitochondrial DNA molecules and their evolutionary relationships. Genetics **97**:125–143.

Cann, R. L., W. M. Brown, and A. C. Wilson. 1982. Evolution of human mitochondrial DNA: a preliminary report. Pp. 156–165 *in* B. Bonné-Tamir, ed. Human genetics. Part A. The unfolding genome. Liss, New York.

Czelusniak, J., M. Goodman, and G. W. Moore. 1978. On investigating the statistical properties of the populous path algorithm by computer simulation: counterconclusions to those of Tateno and Nei. J. Mol. Evol. **11**:75–85.

Estabrook, G. F., C. S. Johnson, Jr., and F. R. McMorris. 1976. A mathematical foundation for the analysis of cladistic character compatibility. Math. Biosci. **29**:181–187.

Farris, J. S. 1970. Methods for computing Wagner trees. Syst. Zool. **19**:83–92.

Farris, J. S. 1977. On the phenetic approach to vertebrate classification. Pp. 823–850 *in* M. K. Hecht, P. C. Goody, and B. M. Hecht, eds. Major patterns in vertebrate evolution. Plenum, New York.

Felsenstein, J. 1978. Cases in which parsimony and compatibility methods will be positively misleading. Syst. Zool. **27**:401–410.

Ferris, S. D., W. M. Brown, W. S. Davidson, and A. C. Wilson. 1981*a*. Extensive polymorphism in the mitochondrial DNA of apes. Proc. Natl. Acad. Sci. USA **78**:6319–6323.

Ferris, S. D., R. D. Sage, C.-M. Huang, J. T. Nielsen, U. Ritte, and A. C. Wilson. 1983*a*. Flow of mitochondrial DNA across a species boundary. Proc. Natl. Acad. Sci. USA **80**:2290–2294.

Ferris, S. D., R. D. Sage, E. M. Prager, U. Ritte, and A. C. Wilson. 1983*b*. Mitochondrial DNA evolution in mice. Genetics **105**:681–721.

Ferris, S. D., A. C. Wilson, and W. M. Brown. 1981*b*. Evolutionary tree for apes and humans based on cleavage maps of mitochondrial DNA. Proc. Natl. Acad. Sci. USA **78:** 2432–2436.

Fitch, W. M., and E. Margoliash. 1967. Construction of phylogenetic trees. Science **155:** 279–284.

Goodman, M., G. W. Moore, J. Barnabas, and G. Matsuda. 1974. The phylogeny of human globin genes investigated by the maximum parsimony method. J. Mol. Evol. **3:**1–48.

Goodman, M., A. E. Romero-Herrera, H. Dene, J. Czelusniak, and R. E. Tashian. 1982. Amino acid sequence evidence on the phylogeny of primates and other eutherians. Pp. 115–187 *in* M. Goodman, ed. Macromolecular sequences in systematic and evolutionary biology. Plenum, New York.

Holmquist, R. 1978. The augmentation algorithm and molecular phylogenetic trees. J. Mol. Evol. **12:**17–24, 369 (erratum).

Kaplan, N., and K. Risko. 1981. An improved method for estimating sequence divergence of DNA using restriction endonuclease mappings. J. Mol. Evol. **17:**156–162.

Kimura, M. 1981*a*. Was globin evolution very rapid in its early stages? a dubious case against the rate-constancy hypothesis. J. Mol. Evol. **17:**110–113.

Kimura, M. 1981*b*. Doubt about studies of globin evolution based on maximum parsimony codons and the augmentation procedure. J. Mol. Evol. **17:**121–122.

Klotz, L. C., and R. L. Blanken. 1981. A practical method for calculating evolutionary trees from sequence data. J. Theor. Biol. **91:**261–272.

Langley, C. M., and W. M. Fitch. 1974. An examination of the constancy of the rate of molecular evolution. J. Mol. Evol. **3:**161–177.

Li, W.-H. 1981*a*. A simulation study of Nei and Li's model for estimating DNA divergence from restriction enzyme maps. J. Mol. Evol. **17:**251–255.

Li, W.-H. 1981*b*. Simple method for constructing phylogenetic trees from distance matrices. Proc. Natl. Acad. Sci. USA **78:**1085–1089.

Nei, M. 1982. Evolution of human races at the gene level. Pp. 167–181 *in* B. Bonné-Tamir, ed. Human genetics. Part A. The unfolding genome. Liss, New York.

Nei, M., and W.-H. Li. 1979. Mathematical model for studying genetic variation in terms of restriction endonucleases. Proc. Natl. Acad. Sci. USA **76:**5269–5273.

Nei, M., J. C. Stephens, and N. Saitou. 1985. Methods for computing the standard errors of branching points in an evolutionary tree and their application to molecular data from humans and apes. Mol. Biol. Evol. **2:**66–85.

Nei, M., and F. Tajima. 1983. Maximum likelihood estimation of the number of nucleotide substitutions from restriction sites data. Genetics **105:**207–217.

Nei, M., F. Tajima, and Y. Tateno. 1983. Accuracy of estimated phylogenetic trees from molecular data. II. Gene frequency data. J. Mol. Evol. **19:**153–170.

Nei, M., and Y. Tateno. 1979. Augmentation algorithm: a reply to Holmquist. J. Mol. Evol. **13:**167–171.

Peacock, D., and D. Boulter. 1975. Use of amino acid sequence data in phylogeny and evaluation of methods using computer simulation. J. Mol. Biol. **95:**513–527.

Sibley, C. G., and J. E. Ahlquist. 1984. The phylogeny of the hominoid primates, as indicated by DNA-DNA hybridization. J. Mol. Evol. **20:**2–15.

Sneath, P. H. A., and R. R. Sokal. 1973. Numerical taxonomy. W. H. Freeman, San Francisco.

Tateno, Y., and M. Nei. 1978. Goodman et al.'s method for augmenting the number of nucleotide substitutions. J. Mol. Evol. **11:**67–73.

Tateno, Y., M. Nei, and F. Tajima. 1982. Accuracy of estimated phylogenetic trees from molecular data. I. Distantly related species. J. Mol. Evol. **18:**387–404.

Templeton, A. R. 1983*a*. Phylogenetic inference from restriction endonuclease cleavage site maps with particular reference to the evolution of humans and the apes. Evolution **37:** 221–244.

TEMPLETON, A. R. 1983*b*. Convergent evolution and nonparametric inferences from restriction data and DNA sequences. Pp. 151–179 *in* B. S. WEIR, ed. Statistical analysis of DNA sequence data. Marcel Dekker, New York and Basel.

YONEKAWA, H., K. MORIWAKI, O. GOTOH, J.-I. HAYASHI, J. WATANABE, N. MIYASHITA, M. L. PETRAS, and Y. TAGASHIRA. 1981. Evolutionary relationships among five subspecies of *Mus musculus* based on restriction enzyme cleavage patterns of mitochondrial DNA. Genetics **98**:801–816.

YUNIS, J. J., and O. PRAKASH. 1982. The origins of man: a chromosomal pictorial legacy. Science **215**:1525–1530.

ROBERT K. SELANDER, reviewing editor

Received June 13, 1984; revision received December 4, 1984.