EVOLUTIONARY CHANGE OF RESTRICTION CLEAVAGE SITES AND PHYLOGENETIC INFERENCE

WEN-HSIUNG LI

Center for Demographic and Population Genetics, University of Texas, Houston, Texas 77030

Manuscript received July 19, 1985
Revised copy accepted January 17, 1986

ABSTRACT

Mathematical formulas are developed for the evolutionary change of restriction cleavage sites in a DNA sequence, allowing unequal rates between transitional and transversional types of nucleotide substitution. Formulas are also developed for the probability of having a particular pattern of site changes among evolutionary lineages, such as parallel gains or losses of sites, and for inferring the presence or absence of a restriction site in an ancestral sequence from data on the present-day sequences. The unordered compatibility method is proposed for inferring the phylogenetic relationships among relatively closely related organisms, treating restriction sites as cladistic characters. Formulas are derived for the probability (P^+) of obtaining the correct network for a given number (N) of informative sites for the cases of four and five species. These formulas are applied to evaluate the performance of the method and to estimate the N value required for P^+ to be 95% or larger. The method performs well when the branches between ancestral nodes and the branches leading to the two most recent species are more or less equal in length, but performs poorly when the latter two branches are considerably longer than the former.

THE restriction enzyme technique has been frequently used in evolutionary studies because it provides a quick means for studying nucleotide variation within and between populations and for studying the phylogenetic relationships among closely related organisms (AVISE, LANSMAN and SHADE 1979; BROWN, GEORGE and WILSON 1979; BROWN and SIMPSON 1981; FERRIS, WILSON and BROWN 1981; Ferris *et al.* 1983; POWELL 1983; CANN, BROWN and WILSON 1984). In view of the importance of this technique, many authors have proposed models for analyzing restriction site data (*e.g.*, UPHOLT 1977; KAPLAN and LANGLEY 1979; NEI and LI 1979; ENGELS 1981; NEI and TAJIMA 1983); however, many problems are still not well explored. One problem is how to infer the presence or absence of a restriction site in an ancestral sequence from observations on present-day sequences. A limited analysis of this problem has recently been conducted by NEI and TAJIMA (1985). I shall study it in more detail. To this end, however, one must know how to evaluate the probability of obtaining a particular pattern of restriction site changes under a given evolutionary tree. NEI and TAJIMA (1985) have studied this probability under

the assumption of random substitution among the four types of nucleotides. I shall extend their results to the case of nonrandom substitution.

Another problem that needs to be explored further is how to reconstruct a phylogenetic tree from restriction site data; this has recently become a controversial issue. TEMPLETON (1983a,b) criticized methods based on NEI and LI's (1979) distance measure and advocated analyzing restriction site data through a combination of Wagner parsimony and character compatibility methods. NEI and TAJIMA (1985) showed that parsimony rules do not hold well under certain conditions and cautioned against uncritical use of the maximum parsimony approach. This methodological dispute has arisen because there is actually no simple answer to the question "What is the best method for reconstructing phylogenetic trees from molecular data?" The "best" method usually varies from situation to situation, and no method is error-free, even in a situation where it is supposed to perform best (*e.g.*, see EDWARDS and CAVALLI-SFORZA 1964; TATENO, NEI and TAJIMA 1982; FELSENTEIN 1984). Therefore, whatever the preferred method may be, one should know its strengths and weaknesses. For this reason, it is important to evaluate the probability of obtaining the correct tree by a particular method under various conditions. I shall show how the theoretical results developed in this paper can be used to evaluate this probability for the character compatibility method (LE QUESNE 1969; ESTABROOK and MCMORRIS 1980). DEBRY and SLADE (1985) have recently studied the probability of obtaining the correct tree by the Wagner parsimony method or the Dollo parsimony method. Their treatment was based on their approximate formulas for the evolutionary change of restriction sites. A more rigorous treatment can be done using the exact formulas given by NEI and TAJIMA (1985) and in this paper, although this will not be pursued in the present study.

BASIC THEORY

Evolutionary change of nucleotides in a DNA sequence: In order to study how restriction sites change with time, one must first study how nucleotides in a DNA sequence change with time. NEI and LI (1979) studied the case in which nucleotide substitution occurs randomly among the four types of nucleotides. Here, I consider the case where the rate of transitional substitution (changes between C and T or between A and G) differs from that of transversional substitution. (A more general formulation can be done, but the mathematics for the evolutionary change of restriction sites become complicated.) This consideration is important, because there is a strong bias favoring transitions in both mitochondrial DNA (BROWN *et al.* 1982; AQUADRO and GREENBERG 1983) and nuclear DNA (FITCH 1967; GOJOBORI, LI and GRAUR 1982; LI, WU and LUO 1984, 1985).

I shall use KIMURA's (1980) two-parameter model of nucleotide substitution. In this model the probabilities of having a transitional change and of having a transversional change per nucleotide site per unit time are α and 2β, respectively, and the total rate of change is $\lambda = \alpha + 2\beta$ per site per unit time. (The transversion rate of 2β comes from the fact that for each nucleotide there are

two possible types of transversional change.) Let $p_A(t)$, $p_T(t)$, $p_C(t)$ and $p_G(t)$ be the probabilities that the nucleotide at a particular site at time t is A, T, C or G, respectively. From formula (2) of AOKI, TATENO and TAKAHATA (1981) one can show that

$$p_A(t) = 1/4 + 1/2[p_A(0) + p_G(0) - 1/2]e^{-4\beta t} + 1/2[p_A(0) - p_G(0)]e^{-2(\alpha+\beta)t}, \quad (1a)$$

$$p_T(t) = 1/4 + 1/2[p_T(0) + p_C(0) - 1/2]e^{-4\beta t} + 1/2[p_T(0) - p_C(0)]e^{-2(\alpha+\beta)t}, \quad (1b)$$

$$p_C(t) = 1/4 + 1/2[p_C(0) + p_T(0) - 1/2]e^{-4\beta t} + 1/2[p_C(0) - p_T(0)]e^{-2(\alpha+\beta)t}, \quad (1c)$$

$$p_G(t) = 1/4 + 1/2[p_G(0) + p_A(0) - 1/2]e^{-4\beta t} + 1/2[p_G(0) - p_A(0)]e^{-2(\alpha+\beta)t}. \quad (1d)$$

Let $p_{ij}(t)$ be the probability that at time t the nucleotide at the site under consideration is j, given that the initial nucleotide was i, where i, j = A, T, C or G. Putting $p_A(0) = 1$ and $p_G(0) = 0$ into (1a), one obtains

$$p(t) = p_{AA}(t) = \tfrac{1}{4} + \tfrac{1}{4}e^{-4\beta t} + \tfrac{1}{2}e^{-2(\alpha+\beta)t}. \quad (2)$$

It can be shown that (2) holds also for $p_{TT}(t)$, $p_{CC}(t)$ and $p_{GG}(t)$. Therefore, $p(t)$ represents the probability that the nucleotide at time t is the same as the initial nucleotide, regardless of the type of the initial nucleotide. This formula appears to be better than formula (5) of TEMPLETON (1983a), which does not hold for large t. The probability $p_{GA}(t)$ that the nucleotide at time t is A, given that the initial nucleotide was G, is given by

$$q(t) = p_{GA}(t) = \tfrac{1}{4} + \tfrac{1}{4}e^{-4\beta t} - \tfrac{1}{2}e^{-2(\alpha+\beta)t} \quad (3)$$

This can readily be obtained by putting $p_A(0) = 0$ and $p_G(0) = 1$ in (1a). By symmetry, we have $p_{AG}(t) = p_{TC}(t) = p_{CT}(t) = p_{GA}(t)$. Thus, $q(t)$ is the probability that the present nucleotide and the initial nucleotide differ by a transition. The probability that the present nucleotide and the initial nucleotide differ by either of the two types of transversion is $2s(t)$, where $s(t)$ denotes the probability that they differ by a particular type of transversion and is given by

$$s(t) = [1 - p(t) - q(t)]/2 = \tfrac{1}{4} - \tfrac{1}{4}e^{-4\beta t}. \quad (4)$$

It is clear from (1) to (4) that, as t approaches ∞, the frequencies of the four types of nucleotide will become 1/4. Further, we also have $p_{ij}(t) = p_{ji}(t)$ for any i and j. This equality means that time is reversible; that is, the probability of starting from type i and changing to type j is the same as going backward from j to i. Time reversibility simplifies the study of nucleotide divergence between two sequences. For example, $q(t)$ can be regarded as the proportion of transitional differences between two sequences that separated $t/2$ time units ago. In fact, KIMURA's (1980) formula for this proportion can be readily obtained by putting $t = 2T$ into (3). Further, as will be seen later, time reversibility greatly simplifies the study of evolutionary changes of restriction sites.

Gain and loss of restriction sites in a DNA sequence: Consider a restriction endonuclease with a particular recognition sequence of r nucleotides, and denote by W_{ij} a sequence of r nucleotides which differ from the recognition sequence by i transitional and j transversional differences. For example, sup-

pose that the recognition sequence is GAATTC, *i.e.*, the *Eco*RI recognition sequence. Then the sequence ACATTC is of the W_{11} type because it differs from the recognition sequence by a transition (position 1) and a transversion (position 2). We shall assume that the equilibrium condition holds so that the four nucleotides are equally frequent and randomly distributed along the DNA sequence under study. Under this assumption, the probability that a randomly chosen sequence of r nucleotides is of the W_{ij} type is given by

$$a_{ij} = \frac{r!}{i!j!(r-i-j)!}\left(\frac{1}{4}\right)^i\left(\frac{2}{4}\right)^j\left(\frac{1}{4}\right)^{r-i-j} = \frac{r!}{i!j!(r-i-j)!}\,2^j a_{00}, \tag{5}$$

where $a_{00} = (1/4)^r$ is the probability that a randomly chosen sequence of length r is the recognition sequence, *i.e.*, W_{00}.

We now consider the probability $v_{ij,kl}$ that a W_{ij} sequence will evolve into a W_{kl} sequence at time t. We start with a simple example. Suppose that the recognition sequence is the *Eco*RI sequence GAATTC and we want to compute the probability that the sequence ACATTC will become the recognition sequence at time t. For this to occur, the nucleotide A at the first position must change to G (a transition), the nucleotide C at the second position must change to A (a transversion) and the last four nucleotides must be the same as the original ones. The probabilities for these three events to occur are $q(t)$, $s(t)$ and $p(t)^4$, respectively, where p, q and s are given by formulas (2) to (4). Therefore, the probability for ACATTC to become the recognition sequence at time t is qsp^4. Since the substitution pattern is the same for the four types of nucleotides, this is the probability for any W_{11} sequence to become W_{00} at time t, *i.e.*, $v_{11,00}$. In general, we can show that

$$v_{ij,00}(t) = q^i s^j p^{r-i-j}. \tag{6}$$

The probability for a W_{00} sequence to become one of the W_{ij} sequences can be obtained by considering the multinomial expansion of $(q + 2s + p)^r$. The problem is to find the term with i q's, j $(2s)$'s and $(r\text{-}i\text{-}j)$ p's. That is,

$$v_{00,ij}(t) = \frac{r!}{i!j!(r-i-j)!}\,q^i(2s)^j p^{r-i-j}. \tag{7}$$

In general, we can show that the probability for a W_{ij} sequence to change to any of the W_{kl} sequences is given by

$$\begin{aligned} v_{ij,kl}(t) = \sum_{j_2}\sum_{j_1}\sum_{i_2}\sum_{i_1} &\frac{i!}{i_1!i_2!(i-i_1-i_2)!}\,\frac{j!}{j_1!j_2!(j-j_1-j_2)!} \\ &\times \frac{(r-i-j)!}{(k-i_1-j_1)!(l-i_2-j_2)!m!}\,2^{l-j_2}p^{i_1+m}s^{j+l-2j_2}\times q^{i+k-2i_1-i_2-j_1}(p+q)^{j_2}, \end{aligned} \tag{8}$$

where $0 \le i_1 \le \min(k, i)$; $0 \le j_1 \le \min(k - i_1, j)$; $0 \le i_2 \le \min(i - i_1, l)$; $0 \le j_2 \le \min(l - i_2, j - j_1)$; $m = r - i - j - k + i_1 + j_1 - l + i_2 + j_2$, and $\min(a, b)$ means the smaller of a and b. The derivation of (8) is tedious, but can be accomplished by considering the expansion of $(p + 2s + q)^i[s + (p + q) + s]^j(q + 2s + p)^{r-i-j}$.

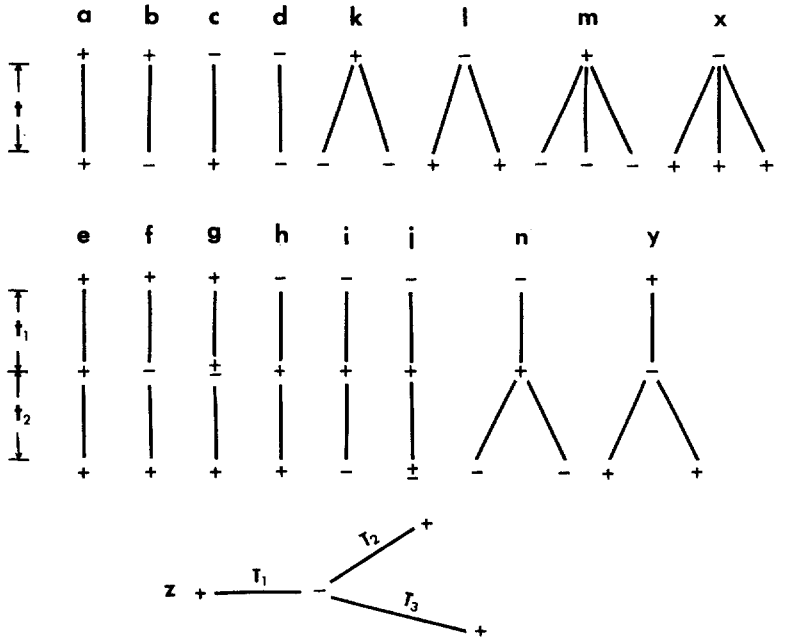

FIGURE 1.—Various types of restriction-site changes: +, the presence of a given restriction site; −, the absence of the restriction site; ±, either the presence or the absence of the restriction site.

From (5), (6) and (7), it is easy to see that $a_{00}v_{00,ij} = a_{ij}v_{ij,00}$. In general, we can show that

$$a_{ij}v_{ij,kl} = a_{kl}v_{kl,ij}. \tag{9}$$

That is, at equilibrium the expected number of W_{ij} sequences changing to W_{kl} sequences during any time period is equal to that of W_{kl}'s changing to W_{ij}'s. A similar relationship has been established by NEI and TAJIMA (1985) for the case of random substitution.

Some basic types of restriction site changes: In Figure 1, cases a–d are the four possible types of restriction site changes in a DNA sequence between two time points. Figure 1a means that the sequence of r nucleotides randomly chosen at time 0 is a restriction site and will also be a restriction site at time t. The former event occurs with probability $a_{00} = (1/4)^r$, whereas the latter occurs with probability $v_{00,00} = p(t)^r$. Therefore, the probability of observing Figure 1a is

$$P_a = a_{00}p(t)^r, \tag{10}$$

where $p(t)$ is given by (2). Similarly, one can show that

$$P_b = P_c = a_{00}[1 - p(t)^r]. \tag{11}$$

$$P_d = 1 - a_{00}[2 - p(t)^r]. \tag{12}$$

Note that $P_a + P_b + P_c + P_d = 1$.

The equality $P_b = P_c$ has been shown to hold for the case of random substitution (NEI and TAJIMA 1985) and should hold for any substitution pattern

under the assumption of equilibrium. Another interesting implication of the equality is that, at equilibrium, time is reversible. Thus, to compute the probability of having a particular pattern of site changes, we may start at the top of the tree and proceed forward or may start at the bottom of the tree and proceed backward.

The above results can be used to derive probabilities for more complicated cases. For the remaining cases in Figure 1, we have

$$P_e = a_{00}p(t_1)^r p(t_2)^r. \tag{13}$$

$$P_g = a_{00}p(t_1 + t_2)^r. \tag{14}$$

$$P_f = P_g - P_e. \tag{15}$$

$$P_h = a_{00}[1 - p(t_1)^r]p(t_2)^r. \tag{16}$$

$$P_i = a_{00}[1 - p(t_1)^r][1 - p(t_2)^r]. \tag{17}$$

$$P_j = a_{00}[1 - p(t_1)^r], \tag{18}$$

$$P_k = a_{00}[1 - p(t)^r]^2. \tag{19}$$

$$P_l = a_{00}[p(2t)^r - p(t)^{2r}]. \tag{20}$$

$$P_m = a_{00}[1 - p(t)^r]^3. \tag{21}$$

$$P_n = a_{00}[1 - p(t_1)^r][1 - p(t_2)^r]^2. \tag{22}$$

$$P_z = a_{00}F(T_1, T_2, T_3), \tag{23}$$

where

$$F(T_1, T_2, T_3) = [q(T_1)q(T_2)q(T_3) + 2s(T_1)s(T_2)s(T_3) + p(T_1)p(T_2)p(T_3)]^r - [p(T_1)p(T_2)p(T_3)]^r. \tag{24}$$

$$P_x = a_{00}F(t, t, t), \tag{25}$$

$$P_y = a_{00}F(t_1, t_2, t_2). \tag{26}$$

Note that Figure 1x is a special case of Figure 1z with $T_1 = T_2 = T_3 = t$ and Figure 1y is a special case of Figure 1z with $T_1 = t_1$ and $T_2 = T_3 = t_2$.

If $t_1 = t_2 = t$, then $P_k = P_i$, $P_l = P_f$, $P_m = P_n$, and $P_x = P_y$. These equalities have been established earlier for the case of random substitution (NEI and TAJIMA 1985) and should hold for any substitution pattern as long as the equilibrium condition obtains.

Table 1 shows the probabilities of several different patterns of evolutionary changes superimposed on known phylogenies. In case 1, $\alpha = \beta = \lambda/3$, and substitution occurs randomly among the four types of nucleotides, whereas in case 2, $\alpha = 0.9\lambda$ and $\beta = 0.05\lambda$, so that transitions occur much more frequently than transversions. It is clear from Table 1 that, for the λt values considered, having an elevated frequency of transitions has little effect on the probability of parallel losses (P_k and P_m) and the probability of gain-loss (P_i and P_n), but increases substantially the probability of parallel gains (P_l and P_x) and the probability of loss-gain (P_f and P_y). TEMPLETON's (1983a, p. 161) conclusion

TABLE 1

Probability (P_i) of having the ith type of restriction site changes given in Figure 1

	λt			
P_i	0.005	0.01	0.05	0.1
(1) $\alpha = \beta = \lambda/3$				
$P_b = P_c$	7.2×10^{-6}	1.4×10^{-5}	6.3×10^{-5}	1.1×10^{-4}
$P_k = P_i$	2.1×10^{-7}	8.3×10^{-7}	1.6×10^{-5}	4.9×10^{-5}
$P_l = P_f$	1.2×10^{-8}	4.4×10^{-8}	7.0×10^{-7}	1.6×10^{-6}
$P_m = P_n$	6.3×10^{-9}	4.8×10^{-8}	4.2×10^{-6}	2.2×10^{-5}
$P_x = P_y$	2.0×10^{-11}	1.4×10^{-10}	8.8×10^{-9}	3.1×10^{-8}
(2) $\alpha = 0.9\lambda$, $\beta = 0.05\lambda$				
$P_b = P_c$	7.2×10^{-6}	1.4×10^{-5}	6.2×10^{-5}	1.1×10^{-4}
$P_k = P_i$	2.1×10^{-7}	8.2×10^{-7}	1.6×10^{-5}	4.7×10^{-5}
$P_l = P_f$	2.8×10^{-8}	1.1×10^{-7}	1.7×10^{-6}	3.8×10^{-6}
$P_m = P_n$	6.3×10^{-9}	4.8×10^{-8}	4.0×10^{-6}	2.0×10^{-5}
$P_x = P_y$	1.2×10^{-10}	8.9×10^{-10}	5.5×10^{-8}	1.9×10^{-7}

It is assumed that $t_1 = t_2 = t$ in Figure 1, and that the number of nucleotides in the recognition sequence is six. The rate of nucleotide substitution is $\lambda = \alpha + 3\beta$ per nucleotide site per unit time. α denotes the rate of transitional substitution per nucleotide site.

that an increase in the frequency of transitions does not have any impact on the probability of parallel (convergent) gains holds only when events involving more than two mutations are negligibly rare. Note further that the probability of parallel gains (P_l) in two lineages or loss-gain (P_f) in one lineage is much smaller than that of parallel losses (P_k) in two lineages or gain-loss (P_i) in one lineage. The same conclusion had been reached earlier by TEMPLETON (1983a). However, as will be seen later, this conclusion may not hold when more than two lineages are involved.

HOW TO INFER THE ANCESTRAL STATUS

We now turn to the problem of how to infer the presence or absence of a restriction site in an ancestral sequence from observations on present-day sequences. This problem is closely related to the problem of parallel evolution (*e.g.*, parallel gains) and also to the problem of inferring the minimum number of restriction site changes. I shall consider only the cases of two or three sequences (species), because the problem becomes complicated when more than three sequences are involved; however, see the results in the next section.

Two species: There are three possible situations. First, the restriction site is present in both species. Such a situation can arise in two ways: (1) the restriction site was already present in the common ancestor, and (2) the site was absent in the common ancestor but has emerged in both species. The probability of the first evolutionary scenario is P_e in (13) with $t_1 = t_2 = t$, where t is the divergence time between the two species. The probability of the second scenario is P_l in (20). Therefore, given the condition that the restriction site is present in both species, the probability that it was present in the ancestral

TABLE 2

Probability (P) for the presence of a restriction site in the common ancestor of two species

α	β	λt	P		
			case 1	case 2	case 3
$\lambda/3$	$\lambda/3$	0.01	1.000	0.486	8.2×10^{-7}
		0.10	0.979	0.360	4.8×10^{-5}
		0.50	0.534	0.062	2.1×10^{-4}
		1.00	0.085	0.008	2.4×10^{-4}
0.9λ	0.05λ	0.01	0.999	0.487	8.2×10^{-7}
		0.10	0.953	0.368	4.7×10^{-5}
		0.50	0.369	0.084	2.0×10^{-4}
		1.00	0.078	0.022	2.3×10^{-4}

The number of nucleotides in the recognition sequence is six. The rate of nucleotide substitution is $\lambda = \alpha + 2\beta$ per nucleotide site per unit time. α denotes the rate of transitional substitution per nucleotide site. In case 1 the restriction site is present in both species, in case 2 it is present in only one of the two species, and in case 3 it is absent in both species.

sequence is

$$P_+(t) = \frac{a_{00}p(t)^{2r}}{a_{00}p(t)^{2r} + a_{00}[p(2t)^r - p(t)^{2r}]} = p(t)^{2r}/p(2t)^r, \tag{27}$$

and the probability that it was absent in the ancestral sequence is $1 - P_+$.

Second, the restriction site is present in only one of the two species. Given this observation, the probability that the restriction site was present in the ancestral sequence can be shown to be equal to

$$P_+(t) = \frac{p(t)^r[1 - p(t)^r]}{1 - p(2t)^r}. \tag{28}$$

Third, the restriction site is absent in both species. In this case,

$$P_+(t) = \frac{a_{00}[1 - p(t)^r]^2}{1 - a_{00}[2 - p(2t)^r]}. \tag{29}$$

Table 2 shows some numerical results. In case 1, the restriction site is present in both species. Obviously, P_+ should be 1 at $t = 0$. As t increases, P_+ decreases, first slowly and then relatively rapidly after $\lambda t = 0.1$. Therefore, if λt is 0.1 or smaller, the sites in the two species are likely to have descended from an ancestral site but, if λt is 1 or larger, the two sites are probably due to parallel gains. A high proportion of transitional substitutions causes a slight reduction in P_+. In case 2, the restriction site is present in only one of the two species. In this case, P_+ is less than 0.5 when $t > 0$ and the rate of decrease in P_+ is faster than in case 1. Therefore, case 2 is more likely to arise from the emergence of a new site in one of the two species than from the loss of an ancestral site in one of the two species. However, unlike case 1, a high proportion of transitional substitutions increases P_+ to some extent. In case 3, the restriction site is absent in both species. P_+ is 0 at $t = 0$ and is virtually negligible for all t values. In all three cases, when t is very large, present observations should have little relevance to the status in the common ancestor, and P_+ should be

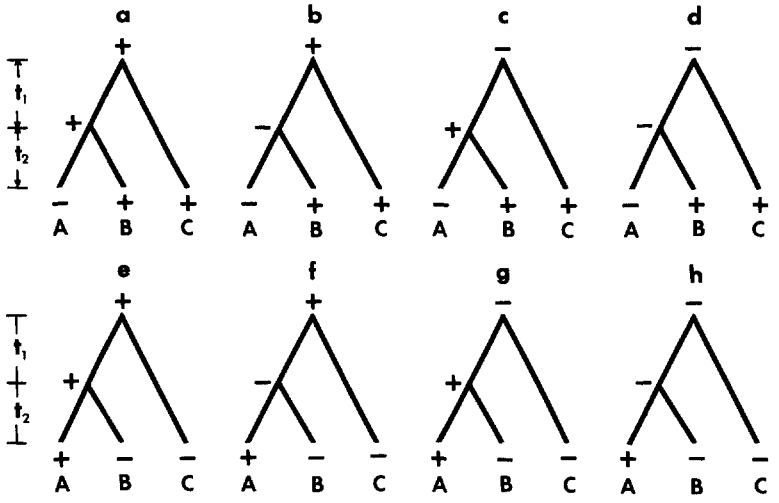

FIGURE 2.—Possible scenarios leading to a given pattern of restriction-site distribution. A, B and C denote three sequences. Scenarios a, b, c, and d are the four possible scenarios that can give rise to the site-distribution pattern shown at the upper row: − in A, but + in B and C. Scenarios e, f, g, and h are the four possible scenarios that can give rise to the pattern at the lower row: + in A but − in B and C.

close to a_{00}, the probability that a randomly chosen sequence of r nucleotides is a restriction site. This can in fact be shown analytically from (27) to (29).

Three species: I consider only two cases: (1) species B and C share a common restriction site not shared with species A, and (2) a restriction site is present in species A but absent in species B and C. In both cases I assume that species A and B are more closely related to each other than either of them is to species C. Figures 2a–d show the four possible situations that can give rise to case 1, while Figures 2e–h show those situations that can give rise to case 2. Figures 2b, d and g are, respectively, the same as Figures 1a, b and d of TEMPLETON (1983a) and represent, respectively, a case of loss-gain, a case of parallel gains and a case of gain-loss. Figure 2e represents a case of parallel losses. TEMPLETON (1983a) was interested in comparing the relative probabilities of observing these four situations. However, instead of studying these situations, he considered the cases of loss-gain and gain-loss in a single lineage (TEMPLETON 1983a, tables 1 and 4) and of parallel gains and parallel losses in two lineages (TEMPLETON 1983a, tables 2 and 3) and then compared their probabilities. A more rigorous comparison is as follows.

The probabilities of having the first four patterns of restriction site changes in Figure 2 can be shown to be

$$P_{(a)} = a_{00}p(t_1)^r[1 - p(t_2)^r]p(t_2)^r p(t_1 + t_2)^r, \tag{30}$$

$$P_{(b)} = a_{00}p(t_1 + t_2)^r[p(t_1 + t_2)^r - p(t_1)^r p(t_2)^r - F(t_1, t_2, t_2)] \tag{31}$$

$$P_{(c)} = a_{00}p(t_2)^r[1 - p(t_2)^r]p(2t_1 + t_2)^r - P_{(a)}, \tag{32}$$

$$P_{(d)} = a_{00}[p(2t_1 + 2t_2)^r - p(t_2)^r p(2t_1 + t_2)^r - F(2t_1 + t_2, t_2, t_2)] - P_{(b)}, \tag{33}$$

TABLE 3

Relative probabilities of having the evolutionary scenarios (a) to (d) in Figure 2 and the probability (P_+) that the restriction site was present in the common ancester of the three species

α	λt_1	λt_2	$P'_{(a)}$	$P'_{(b)}$	$P'_{(c)}$	$P'_{(d)}$	$P_+ = P'_{(a)} + P'_{(b)}$
$\lambda/3$	0.01	0.01	0.9893	0.0034	0.0004	0.0069	0.993
		0.10	0.9436	0.0043	0.0022	0.0499	0.948
	0.05	0.01	0.9563	0.0167	0.0060	0.0210	0.973
		0.10	0.8958	0.0208	0.0144	0.0690	0.917
	0.10	0.01	0.9067	0.0322	0.0216	0.0395	0.939
		0.10	0.8296	0.0393	0.0372	0.0939	0.869
0.9λ	0.01	0.01	0.9746	0.0081	0.0010	0.0163	0.983
		0.10	0.8802	0.0093	0.0047	0.1057	0.890
	0.05	0.01	0.9025	0.0378	0.0133	0.0465	0.940
		0.10	0.7910	0.0423	0.0293	0.1374	0.833
	0.10	0.01	0.8072	0.0675	0.0442	0.0811	0.875
		0.10	0.6824	0.0737	0.0687	0.1752	0.756

The number of nucleotides in the recognition sequence is six. The rate of nucleotide substitution is λ per nucleotide site per unit time. α denotes the rate of transitional substitution per nucleotide site. The relative probabilities are defined as $P'_{(a)} = P_{(a)}/S$, $P'_{(b)} = P_{(b)}/S$, $P'_{(c)} = P_{(c)}/S$ and $P'_{(d)} = P_{(d)}/S$, where $S = P_{(a)} + P_{(b)} + P_{(c)} + P_{(d)}$.

in which $F(\cdot, \cdot, \cdot)$ is defined by (24).

Table 3 shows some numerical results for the relative probabilities of the four scenarios. Scenario (a) would be the most likely, because it requires only a single mutational event, *i.e.*, loss of the restriction site in species A (TEMPLETON 1983a). This is indeed supported by the numerical results. Note, however, that if λt_1 or λt_2 is 0.05 or larger, scenario (a) is not the correct explanation for a substantial fraction (larger than 10%) of the cases, particularly if transition is the predominant type of mutation. The next most likely scenario is (d), which requires (at least) two mutational events, a gain in both species B and C. When $\lambda t_1 = \lambda t_2 = 0.1$, this scenario can account for 9% of the cases if $\alpha = \lambda/3$ and for 18% of the cases if $\alpha = 0.9\lambda$. Like scenario (d), scenario (b) also requires two mutational events, a loss before the A–B split and a gain in species B. This scenario is somewhat less likely than scenario (d), particularly when transition is the predominant type of mutation. Scenario (c) requires at least three mutational events and is the least likely scenario among the four. The P_+ value is high when λt_1 and λt_2 are small. Thus, if the three species are closely related, the observed pattern of restriction sites among the three species can be taken as a good indication that the common ancestral sequence was a recognition sequence. This inference is, however, in error in more than 10% of the cases, if λt_1 and λt_2 are equal to or larger than 0.1.

The probabilities of having the last four patterns of restriction site changes in Figure 2 can be shown to be

$$P_{(e)} = a_{00}p(t_1)^r p(t_2)^r[1 - p(t_2)^r][1 - p(t_1 + t_2)^r], \quad (34)$$

$$P_{(f)} = a_{00}[p(t_1 + t_2)^r - p(t_1)^r p(t_2)^r - F(t_1, t_2, t_2)][1 - p(t_1 + t_2)^r], \quad (35)$$

TABLE 4

Relative probabilities of having the evolutionary scenarios (e) to (h) in Figure 2 and the probability (P_+) that the restriction site was present in the common ancestor of the three species

α	λt_1	λt_2	$P'_{(e)}$	$P'_{(f)}$	$P'_{(g)}$	$P'_{(h)}$	$P_+ = P'_{(e)} + P'_{(f)}$
$\lambda/3$	0.01	0.01	0.0875	0.0728	0.0476	0.7921	0.160
		0.10	0.1992	0.1582	0.0253	0.6173	0.357
	0.05	0.01	0.1470	0.1245	0.1677	0.5608	0.271
		0.10	0.1783	0.1450	0.1038	0.5729	0.323
	0.10	0.01	0.1520	0.1315	0.2523	0.4642	0.283
		0.10	0.1482	0.1241	0.1705	0.5572	0.272
0.9λ	0.01	0.01	0.0884	0.0741	0.0478	0.7897	0.162
		0.10	0.205	0.1647	0.0258	0.6041	0.370
	0.05	0.01	0.1499	0.1306	0.1677	0.5518	0.280
		0.10	0.1841	0.1555	0.1051	0.5553	0.340
	0.10	0.01	0.1563	0.1427	0.2503	0.4507	0.299
		0.10	0.1539	0.1384	0.1716	0.5361	0.292

The number of nucleotides in the recognition sequence is six. The rate of nucleotide substitution is λ per nucleotide site per unit time. α denotes the rate of transitional substitution per nucleotide site. The relative probabilities are defined as $P'_{(e)} = P_{(e)}/S$, $P'_{(f)} = P_{(f)}/S$, $P'_{(g)} = P_{(g)}/S$ and $P'_{(h)} = P_{(h)}$, where $S = P_{(e)} + P_{(f)} + P_{(g)} + P_{(h)}$.

$$P_{(g)} = a_{00}p(t_2)^r[1 - p(t_2)^r][1 - p(2t_1 + t_2)^r] - P_{(e)}, \tag{36}$$

$$P_{(h)} = a_{00}[1 - p(t_2)^r - p(2t_1 + 2t_2)^r + p(t_2)^r p(2t_1 + t_2)^r - p(2t_2)^r + p(t_2)^{2r} + F(2t_1 + t_2, t_2, t_2)] - P_{(f)}. \tag{37}$$

Table 4 shows some numerical results for the relative probabilities of the last four scenarios. Intuitively, scenario (h) should be the most likely, because it requires only a single mutational event, *i.e.*, a gain in species A. This is indeed true for all of the λt_1 and λt_2 values used. However, the $P'_{(h)}$ value in general decreases as λt_1 increases, and it becomes only 50% or less when λt_1 is 0.10. In the majority of cases, scenario (e) is the second most likely scenario. In this scenario the three species shared a common ancestral restriction site, but the site became lost in species B and C (*i.e.*, parallel losses). Scenario (g) requires a gain and a loss. When λt_1 is small, say 0.01, the probability of gaining a new site is small, and thus, scenario (g) is the least likely among the four. It is, however, the second most likely scenario when λt_1 is 0.10. We note that, although the restriction site is absent in two of the three species, the probability (P_+) of its presence in the common ancestor is at least 15% and can be as large as 37% for the λt_i values considered in Table 4. P_+ first increases as λt_1 or λt_2 increases, but decreases as λt_1 or λt_2 becomes relatively large.

Scenario (f) requires two losses and a gain. It is seen that $P'_{(f)}$ is the same order of magnitude as either $P'_{(e)}$ or $P'_{(g)}$. This contradicts TEMPLETON's (1983a) conclusion that two parallel losses (Figure 2e) and gain-loss (Figure 2g) are far more probable, by one order of magnitude, than loss-gain (Figure 2f). His

conclusion was based on the probabilities of gain-loss and loss-gain in a single evolutionary lineage without considering the status of other lineages. The present numerical results show that it is important to consider the status of other related lineages and also the magnitudes of λt_1 and λt_2. Indeed, if λt_1 is small and λt_2 is large, then even parallel gains (Figure 2d) can be more probable than gain-loss (Figure 2g). For example, if $\lambda t_1 = 0.005$, $\lambda t_2 = 0.05$, and $\alpha = 0.9\lambda$, then $P_{(d)} = 1.7 \times 10^{-6}$, whereas $P_{(g)} = 1.3 \times 10^{-6}$. This is again contradictory to Templeton's (1983a) conclusion that a gain-loss (Figure 2g) is far more probable than parallel gains (Figure 2d). It is therefore dangerous to draw a conclusion about the evolutionary pattern of site gains and losses without any knowledge about the divergence times among species, particularly if many species are involved.

The numerical results in Table 4 show that the relative probabilities are complicated functions of λt_1 and λt_2, but are not much affected by the α value, a situation very different from those in Table 3. As a consequence, the P_+ value in Table 4 is also not much affected by the α value. The P_+ value in Table 3 always decreases as λt_1 or λt_2 increases, whereas the P_+ value in Table 4 may sometimes increase as λt_2 increases.

PHYLOGENETIC INFERENCE

The method

The method to be used is the unordered compatibility method (LE QUESNE 1969; ESTABROOK and MCMORRIS 1980), which gives unrooted trees or networks. TEMPLETON (1983b) proposed to use this method in combination with the Wagner parsimony method. In his algorithm, all the sites cut by a particular enzyme are pooled, and a Wagner parsimony tree is obtained for each enzyme used. The compatibility method is then used to obtain a network that is most compatible with enzyme-specific trees. DEBRY and SLADE (1985) pointed out that a site-by-site analysis is superior because the sites recognized by the same enzyme may not evolve at the same rate (see also, TEMPLETON 1986). If each recognition site is treated as a cladistic character, Templeton's algorithm becomes identical to the unordered compatibility method.

The method is illustrated below, using FERRIS, WILSON and BROWN's (1981) restriction site data on mitochondrial DNAs from human (H), chimpanzee (C), gorilla (G), orangutan (O) and gibbon (Gi). In the compatibility approach, a restriction site is not informative for discriminating alternative topologies if it is present in all species, all but one species or only one species, and such sites are ignored. For the informative sites, a total of 15 patterns of restriction site distributions are observed. They are presented in Table 5 in decreasing frequency order.

In the compatibility approach we search for the largest set of characters that are compatible with the same network and choose that network as the result. First, the network (CG)-H-(OGi) (Figure 3a) is compatible with patterns 1, 2, 10 and 14; thus, in total, it is compatible with 12 informative restriction sites (= 6 + 4 + 1 + 1). Second, the network (CG)-(HO)-Gi (Figure 3b) is compatible

TABLE 5

Compatibility analysis of restriction site distributions among human (H), chimpanzee (C), gorilla (G), orangutan (O) and gibbon (Gi)

Patterns	Species					Restriction sites[a]	No. of sites	Sign test[b]	
	C	G	H	O	Gi			1	2
1	+	+	−	−	−	17x, 32g, 51j, 60k, 86c, 95h	6	+	+
2	−	−	−	+	+	35o, 38x, 66z, 81y	4	0	0
3	+	−	+	−	−	41z, 45z, 50x	3	−	−
4	+	+	−	−	+	20f, 70h	2	+	0
5	+	−	+	+	−	15f, 65y	2	−	0
6	+	−	−	−	+	33w, 95x	2	0	0
7	−	−	+	−	+	23m, 61w	2	+	0
8	−	−	+	+	−	10l, 21o	2	+	0
9	−	+	+	−	−	19o, 55x	2	0	0
10	+	+	+	−	−	29x	1	0	0
11	−	+	−	+	−	52i	1	−	0
12	+	−	−	+	−	6m	1	0	0
13	−	+	−	−	+	47w	1	−	0
14	−	−	+	+	+	95o	1	+	+
15	−	+	+	−	+	85h	1	0	0
No. of − signs								7	3
Total no. of − and + signs								20	10

[a] The data are from table 1 of FERRIS, WILSON and BROWN (1981). A restriction site that appears in only one species, four species or all five species is not informative for compatibility analysis and is not included in the table. Each site is designated by a number for its position in the map and by a letter for the enzyme used. The 86c and 52i sites (c = *Hpa*I, i = *Sal*I) are also recognized by enzyme o (*Hin*cII), which recognizes four sequences (see text). To avoid overlap, the data from enzyme o for these two sites are excluded. For the same reason, the data from enzyme m (*Ava*I) for the 10l site (l = *Xho*I) are also excluded.

[b] The comparison is between the two networks (CG)-H-(OGi) and (HC)-G-(OGi). See text for the two tests.

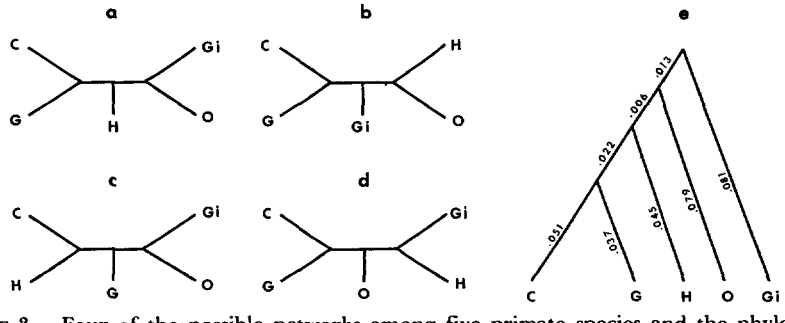

FIGURE 3.—Four of the possible networks among five primate species and the phylogeny inferred from the restriction site data. The five species are chimpanzee (C), gorilla (G), human (H), orangutan (O) and gibbon (Gi). The numbers on the branches of the phylogenetic tree are the estimated branch lengths (numbers of nucleotide substitutions per nucleotide site).

with patterns 1, 4, 8 and 14; in total, it is compatible with 11 informative sites. Third, the network (CG)-O-(HGi) (Figure 3d) is compatible with patterns 1, 7 and 14, *i.e.*, with 9 informative sites. Fourth, the network (HC)-G-(OGi) (Figure 3c) is compatible with patterns 2, 3 and 10, *i.e.*, with 8 informative sites. It can be shown that all other alternative networks are each compatible with fewer than 8 informative sites. Thus, network (a), (CG)-H-(OGi), is the best choice. It is also the most parsimonious in terms of the number of mutational changes (FERRIS, WILSON and BROWN 1981).

A network may be converted into a phylogeny using information from outside sources, such as morphological and paleontological data. The outside information strongly indicates that gibbon branched off earlier than the other four species. Therefore, network (a) is converted into the phylogeny shown in Figure 3. Alternatively, one may first calculate the evolutionary distance between each species pair using NEI and LI's (1979) method or a similar method and then place the root at a place where it produces the maximum separation in terms of distance between the two groups of species on the opposite sides of the root. Applying this method to table 4 of NEI, STEPHENS and SAITOU (1985), which shows the number of nucleotide substitutions per nucleotide site estimated from FERRIS, WILSON and BROWN's (1981) data, one obtains the same phylogeny as that shown in Figure 3.

One way to calculate the branch lengths of a phylogenetic tree is to apply the maximum parsimony principle to each restriction site to infer the minimum number of mutational changes (see FERRIS, WILSON and BROWN 1981). This approach is suitable only if the λt_i values are small; otherwise, it tends to give erroneous estimates (NEI and TAJIMA 1985). A better approach is to use the computational procedure of FITCH and MARGOLIASH (1967), adding the constraint of nonnegative solutions (LI 1981). This procedure is applied to table 4 of NEI, STEPHENS and SAITOU (1985) to obtain the branch lengths shown in Figure 3. Note that in this approach one can at the same time infer the tree root without using any outside information (see above).

It is interesting to know whether network (a) is significantly better than networks (b), (c) and (d) in Figure 3. For this purpose, TEMPLETON (1983b) proposed to use the Wilcoxon signed-rank test. I propose to use the sign test (see also, FELSENSTEIN 1985). Let us compare networks (a) and (c), which differ only in that, in network (a), C and G are clustered together, whereas in network (c), C and H are clustered together. Therefore, for each informative site we may assign the sign + [network (a) favored] if the character state (+ or −) in C is the same as that in G but different from that in H, and the sign − [network (c) favored] if the character state in C is the same as that in H but different from that in G, and 0 (a tie) in all other cases. According to this rule, there are 13 informative sites with the + sign and 7 with the − sign (Table 5). The difference is not significant (see table A8 in SNEDECOR and COCHRAN 1967). This approach is called "test 1" in Table 5. Alternatively, we may assign the sign + if the informative site is compatible with network (a) but not with network (c), − if it is compatible with network (c) but not with network (a), and 0 in all other cases. This is "test 2," and as in "test 1," the

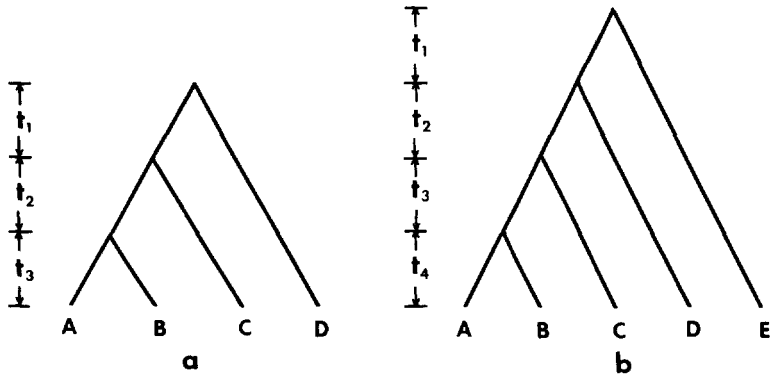

FIGURE 4.—Two-model trees used for studying the performance of the unordered compatibility method.

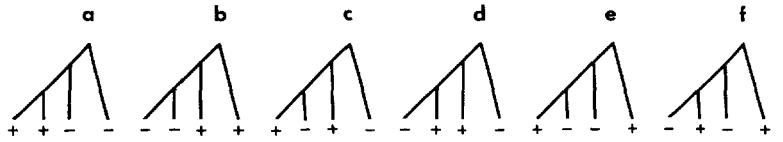

FIGURE 5.—Six possible site-distribution patterns for two +'s and two −'s.

difference between the two networks is nonsignificant. The same conclusion holds for the comparisons between (a) and (b) and between (a) and (d).

Probability of obtaining the correct network

DEBRY and SLADE (1985) have recently studied the probability that a restriction site will evolve in a pattern that results in the correct phylogenetic hypothesis according to the Wagner parsimony or the Dollo parsimony methods. Here, I study this probability for the unordered compatibility method, which is applicable only when the number of species studied is larger than three. I consider two cases: four and five species, using the model trees shown in Figure 4.

Four species: In this case a restriction site is phylogenetically informative only if it is present in two of the four species and absent in the other two species. Figure 5 shows the six possible patterns of site distributions among four species when there are two +'s and two −'s. Only patterns (a) and (b) give the correct network.

The probability (P_a) of obtaining pattern (a) can be calculated as follows. As shown in the upper row of Figure 6,

$$P_a = P_{a1} - P_{a2} - P_{a3} + P^*. \tag{38}$$

P_{a1}, P_{a2} and P_{a3} can easily be obtained by using the results obtained earlier and are given by

$$P_{a1} = a_{00}p(2t_3)^r,$$

$$P_{a2} = a_{00}p(t_3)^{2r}p(2t_2 + t_3)^r + a_{00}F(2t_2 + t_3, t_3, t_3),$$

$$P_{a3} = a_{00}p(t_3)^{2r}p(2t_1 + 2t_2 + t_3)^r + a_{00}F(2t_1 + 2t_2 + t_3, t_3, t_3),$$

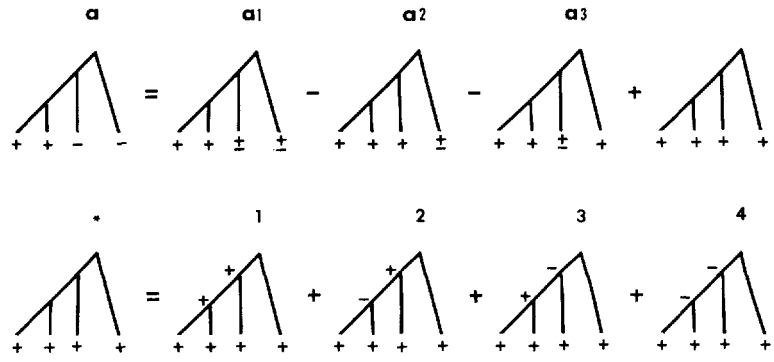

FIGURE 6.—Algebra of site-distribution patterns. *Upper row*: The probability (P_a) of obtaining pattern a is equal to $P_{a1} - P_{a2} - P_{a3} + P^*$, which are the probabilities of obtaining patterns a1, a2, a3, and *, respectively. In evaluating the latter probabilities, all lineages that end with the ± sign can be neglected. *Lower row*: The four scenarios on the right are the four possible scenarios that can give rise to pattern * and $P^* = P_1^* + P_2^* + P_3^* + P_4^*$.

where $F(\cdot, \cdot, \cdot)$ is defined by (24), and the t_i values are those in Figure 4. P^* can be obtained by considering the four possible scenarios shown at the bottom row in Figure 6. Using the results obtained earlier, we can show that the probabilities of having the first three scenarios are

$$P_1^* = a_{00}p(t_3)^{2r}p(t_2)^r p(t_2 + t_3)^r p(2t_1 + t_2 + t_3)^r,$$

$$P_2^* = a_{00}F(t_2, t_3, t_3)p(t_2 + t_3)^r p(2t_1 + t_2 + t_3)^r,$$

$$P_3^* = a_{00}p(t_3)^{2r}F(t_2, t_2 + t_3, 2t_1 + t_2 + t_3).$$

P_4^* is given by

$$P_4^* = a_{00} \sum_j \sum_i v_{00,ij}(2t_1 + t_2 + t_3)v_{ij,00}(t_2 + t_3) \sum_l \sum_k v_{ij,kl}(t_2)v_{kl,00}(t_3)^2,$$

in which $1 \le i + j \le r$ and $1 \le l + k \le r$. When λt_2 is small, this formula can be approximated by

$$\begin{aligned} P_4^* &= a_{00} \sum_j \sum_i v_{00,ij}(2t_1 + t_2 + t_3)v_{ij,00}(t_2 + t_3)v_{ij,00}(t_3)^2 \\ &= a_{00}[\{q(2t_1 + t_2 + t_3)q(t_2 + t_3)q(t_3)^2 + 2s(2t_1 + t_2 + t_3)s(t_2 + t_3)s(t_3)^2 \\ &\quad + p(2t_1 + t_2 + t_3)p(t_2 + t_3)p(t_3)^2\}^r - \{p(2t_1 + t_2 + t_3)p(t_2 + t_3)p(t_3)^2\}^r]. \end{aligned}$$

It can be shown that P_4^* is at least one order of magnitude smaller than P_1^* because it requires at least three parallel gains. Therefore, P_4^* can actually be neglected when computing P^*. At any rate.

$$P^* = P_1^* + P_2^* + P_3^* + P_4^*. \tag{39}$$

In the same manner, we can show that

$$\begin{aligned} P_b = a_{00}[H(t_1, t_2 + t_3) - p(2t_1 + t_2 + t_3)^r p(t_2 + t_3)^{2r} \\ -F(2t_1 + t_2 + t_3, t_2 + t_3, t_2 + t_3)] + P^*, \end{aligned} \tag{40}$$

$$P_c = P_d = P_b + a_{00}[H(t_2, t_3) - H(t_1, t_2 + t_3)], \quad (41)$$

$$P_e = P_f = P_b + a_{00}[H(t_1 + t_2, t_3) - H(t_1, t_2 + t_3)], \quad (42)$$

where

$$H(x, y) = p(2x + 2y)^r - p(2x + y)^r p(y)^{2r} - F(2x + y, y, y). \quad (43)$$

We noted above that only patterns (a) and (b) give the correct network. Therefore, for a single informative site, the probability of obtaining the correct network is equal to $P = (P_a + P_b)/P_I$, where $P_I = P_a + P_b + P_c + P_d + P_e + P_f$ is the total probability of having any of the six-site distribution patterns shown in Figure 5.

Table 6 shows the probabilities of obtaining the site distribution patterns shown in Figure 5. Pattern (a) (++−−) is always the most frequent among the six patterns, particularly when the λt_i values are relatively large. Pattern (b) (−−++) is the second most frequent in the majority of cases, but is less frequent than patterns (c) and (d) (+−+− and −++−) if λt_2 is considerably smaller than both λt_1 and λt_3. Patterns (e) and (f) (+−−+ and −+−+) are always the least frequent among the six patterns.

Table 6 also shows the probability (P) of obtaining the correct network, using one informative site. Interestingly, P depends more on the relative values of λt_2 and λt_3 than on the absolute value of λt_2. Intuitively, P would be low if λt_2 were small, because the chance for a change to occur between the ancestral node of species A, B and C and that of species A and B would be small. However, P is actually high when both λt_2 and λt_3 are small. The reason for this is that P is a relative probability—although P_a and P_b are small when λt_2 and λt_3 are small, P_c, P_d, P_e and P_f are even smaller. At any rate, P is relatively high when λt_2 and λt_3 are of the same order of magnitude, but decreases as λt_3 becomes larger than λt_2. Note further that P decreases as λt_1 increases. Therefore, if our aim is to resolve the phylogenetic relationships among species A, B and C, the outside (fourth) species should be as close to the three species as possible. This agrees with intuition. Table 6 shows that nonrandom substitution causes a slight reduction in P, usually 1–2%, but increases slightly the probability P_I that a randomly chosen sequence of r nucleotides is an informative site.

When there are multiple informative sites, the probability of obtaining the correct network can be calculated as follows. There are only three possible networks: (AB)-(CD), (AC)-(BD) and (AD)-(BC). For each single informative site, the three networks occur with probabilities P, $(1 - P)/2$, and $(1 - P)/2$, respectively. For N informative sites, the networks are distributed according to the following multinomial expansion.

$$[P + (1 - P)/2 + (1 - P)/2]^N = \sum_{m_1} \sum_{m_2} \binom{N}{m_1} \binom{N - m_1}{m_2} P^{m_1} \left(\frac{1 - P}{2}\right)^{N - m_2}, \quad (44)$$

where $0 \leq m_1, m_2 \leq N$ and $m_1 + m_2 \leq N$. Hence, the probability (P^+) of obtaining the correct network is equal to the sum of the terms with m_1 larger

TABLE 6

Probabilities (× 10^5) of obtaining the site distribution patterns of (a–f) shown in Figure 5 and probability (P) of obtaining the correct network by the compatibility method

λt_1	λt_2	λt_3	P_a	P_b	$P_c = P_d$	$P_e = P_f$	P_I	P
0.010	0.010	0.010	1.73	1.07	0.25	0.13	3.56	0.79
			(1.70)	(1.10)	(0.26)	(0.14)	(3.60)	(0.78)
		0.025	1.95	1.13	0.65	0.41	5.20	0.59
		0.050	2.26	1.37	1.24	0.92	7.95	0.46
			(2.29)	(1.49)	(1.33)	(1.03)	(8.50)	(0.44)
	0.025	0.025	3.83	1.92	0.68	0.48	8.07	0.71
			(3.73)	(2.01)	(0.73)	(0.55)	(8.30)	(0.69)
		0.050	3.66	1.85	1.20	0.93	9.77	0.56
	0.050	0.050	5.67	2.18	1.09	0.89	11.81	0.66
			(5.51)	(2.39)	(1.23)	(1.04)	(12.44)	(0.63)
0.025	0.010	0.025	2.32	0.98	0.91	0.37	5.86	0.56
			(2.28)	(1.07)	(0.94)	(0.44)	(6.11)	(0.55)
		0.050	2.64	1.19	1.55	0.80	8.53	0.45
	0.025	0.025	4.25	1.65	0.88	0.42	8.50	0.69
			(4.12)	(1.77)	(0.92)	(0.50)	(8.73)	(0.67)
		0.050	4.03	1.59	1.44	0.81	10.12	0.56
0.050	0.010	0.010	2.39	0.71	0.58	0.10	4.46	0.70
		0.025	2.80	0.77	1.25	0.30	6.67	0.53
			(2.74)	(0.89)	(1.27)	(0.40)	(6.97)	(0.52)
		0.050	3.13	0.94	1.95	0.64	9.25	0.44
	0.025	0.025	4.80	1.28	1.14	0.34	9.04	0.67
			(4.63)	(1.44)	(1.18)	(0.44)	(9.31)	(0.65)
		0.050	4.53	1.25	1.75	0.65	10.58	0.55
	0.050	0.050	6.46	1.46	1.44	0.61	12.02	0.66
			(6.26)	(1.74)	(1.58)	(0.80)	(12.76)	(0.63)

It is assumed that $r = 6$ and that nucleotide substitution occurs randomly, *i.e.*, $\alpha = \lambda/3$, except for those values in parentheses, where $\alpha = 0.9\lambda$ $P_I = P_a + P_b + P_c + P_d + P_e + P_f$ is the probability that a randomly chosen sequence of r nucleotides is an informative site.

than both m_2 and $m_3 = N - m_1 - m_2$. (I neglect the case where a tie for the most compatible network occurs between networks 1 and 2 or networks 1 and 3, *i.e.*, $m_1 = m_2 \geq m_3$ or $m_1 = m_3 \geq m_2$.) For P^+ to be 95% or larger, it requires only 11 or fewer informative sites if the P value in (44) is 0.70 or larger (Table 7). In the case of $\lambda t_1 = \lambda t_2 = \lambda t_3 = 0.025$, $P = 0.69$ and $P_I = 8.5 \times 10^{-5}$ (Table 6). In mammals, the mitochondrial DNA (mtDNA) is about 16,500 nucleotides long (ANDERSON *et al.* 1981). Therefore, when a six-base restriction enzyme is applied to the mtDNAs from four species with the above λt_i values, the expected number of informative sites is $16{,}500 \times 8.5 \times 10^{-5} = 1.4$. (I assume that the four types of nucleotides are equally frequent and randomly distributed.) To have 11 informative sites, one requires about eight (= 11/1.4) six-base enzymes. If $P = 0.60$ and $P_I = 5.0 \times 10^{-5}$ as approximately in the

TABLE 7

Probability (P^+) of obtaining the correct network for a given number (N) of informative sites

	P^+						
P	$N = 5$	$N = 10$	$N = 20$	$N = 30$	$N = 50$	$N = 120$	$N = 250$
0.40	0.32	0.40	0.50	0.58	0.66	0.83	0.93
0.50	0.50	0.63	0.81	0.89	0.96	1.00	1.00
0.60	0.68	0.82	0.96	0.99	1.00	1.00	1.00
0.70	0.84	0.94	1.00	1.00	1.00	1.00	1.00

For a single informative site, networks 1, 2 and 3 occur with probabilities P, $(1 - P)/2$ and $(1 - P)/2$, respectively. The probability in the table refers to the probability that network 1 is strictly more frequent than both networks 2 and 3 in a sample of N informative sites.

case of $\lambda t_1 = \lambda t_2 = 0.01$ and $\lambda t_3 = 0.025$ (Table 6), then for P^+ to be 95% or larger, one requires 20 informative sites or 24 six-base restriction enzymes. This is a rather large number. If $P = 0.50$ and $P_I = 6.0 \times 10^{-5}$, one requires 50 informative sites or 50 six-base enzymes. If P is 0.4 or smaller, the number of informative sites required is far too large (Table 7). Of course, the number required is considerably smaller if one requires P^+ to be 80% rather than 95%.

From Tables 6 and 7 we may conclude that the unordered compatibility method is useful for inferring the phylogenetic relationships among four species if λt_2 and λt_3 are of similar magnitude and if λt_1, λt_2 and λt_3 are of the order of 0.05 or smaller. (The method is also expected to perform well if λt_3 is smaller than λt_2.) The method is, however, not very useful if λt_2 is considerably smaller than λt_3.

Five species: In this case a restriction site is phylogenetically informative only if it is present in two or three of the five species. In either case, there are ten possible site distribution patterns, so the total number of distribution patterns is 20 (Table 8). In order to evaluate the performance of the compatibility method, we need to compute the probability of obtaining each of these 20 patterns. Fortunately, six pattern pairs (*e.g.*, +−+−+ and −+−+−) occur with equal probabilities, so that we need compute the probabilities for only 14 patterns (Table 8).

I now show how to obtain the probabilities for the first two patterns in Table 8; the probabilities for the other patterns can be obtained in the same manner. In the second pattern, ++−−−, the restriction site is present in A and B but absent in C, D and E. We consider this pattern together with another pattern in which the restriction site is present not only in A and B but also in one of the other three species. Let us, for example, take the first pattern, +++−−. We note that (++−−−) + (+++−−) = ++±−−, which is equivalent to pattern (a) in the case of four species (Figure 5a). Therefore, the probability of observing either ++−−− or +++−− can be computed by using (37) with a modification of the λt_i values. After this, the probability of observing ++−−− can be readily obtained if we know the probability of observing +++−−.

The latter probability can be computed as follows. We note that +++−− =

TABLE 8

Relative frequencies of 20 different site distribution patterns among five species

λt_1	λt_2	λt_3	λt_4	1 +++−−	2 ++−−−	3 −−−++	4 −−+++	5 ++−+−	6 −−++−	7 ++−−+	8, 9 +−+−+ −++−+	10, 11 +−++− −+++−	12 −−+−+	13, 14 +−+−− −++−−	15, 16 +−−++ −+−++	17, 18 +−−+− −+−+−	19, 20 +−−−+ −+−−+
0.010	0.010	0.010	0.010	0.234	0.193	0.135	0.104	0.064	0.039	0.037	0.017	0.030	0.025	0.016	0.011	0.005	0.004
			0.025	0.172	0.141	0.094	0.072	0.073	0.038	0.049	0.033	0.050	0.028	0.031	0.025	0.015	0.012
			0.050	0.117	0.106	0.067	0.052	0.069	0.043	0.051	0.041	0.055	0.034	0.046	0.034	0.030	0.024
		0.025	0.025	0.137	0.221	0.079	0.082	0.074	0.055	0.054	0.024	0.034	0.042	0.025	0.020	0.015	0.012
			0.050	0.103	0.168	0.066	0.056	0.068	0.055	0.052	0.032	0.041	0.045	0.041	0.027	0.029	0.024
	0.025	0.010	0.025	0.246	0.135	0.128	0.044	0.057	0.031	0.041	0.028	0.039	0.024	0.042	0.016	0.012	0.010
			0.050	0.160	0.118	0.090	0.036	0.058	0.037	0.045	0.036	0.046	0.030	0.060	0.024	0.026	0.021
		0.025	0.025	0.194	0.217	0.102	0.056	0.062	0.047	0.047	0.021	0.028	0.037	0.034	0.013	0.013	0.011
			0.050	0.138	0.179	0.081	0.041	0.059	0.049	0.047	0.028	0.036	0.040	0.053	0.020	0.025	0.022
0.025	0.010	0.010	0.025	0.190	0.137	0.075	0.056	0.091	0.046	0.039	0.026	0.063	0.024	0.033	0.020	0.018	0.011
			0.050	0.130	0.109	0.057	0.042	0.081	0.049	0.041	0.033	0.065	0.030	0.050	0.028	0.033	0.021
		0.025	0.025	0.153	0.219	0.066	0.066	0.089	0.065	0.044	0.019	0.041	0.036	0.027	0.016	0.017	0.011
			0.050	0.115	0.171	0.057	0.046	0.078	0.062	0.044	0.026	0.048	0.039	0.045	0.022	0.032	0.021
	0.025	0.010	0.025	0.262	0.135	0.106	0.035	0.069	0.036	0.034	0.023	0.047	0.021	0.044	0.013	0.014	0.009
			0.050	0.172	0.122	0.076	0.030	0.067	0.042	0.037	0.029	0.054	0.027	0.064	0.020	0.029	0.019
		0.025	0.050	0.149	0.184	0.070	0.034	0.067	0.054	0.039	0.023	0.041	0.035	0.056	0.017	0.028	0.019
	0.050	0.010	0.050	0.225	0.139	0.085	0.018	0.051	0.033	0.031	0.024	0.041	0.022	0.082	0.012	0.023	0.016

It is assumed that $r = 6$ and that nucleotide substitution occurs randomly.

(+++±±) − (+++++) − (+++±+) + (+++++). The first pattern on the right side involves only three species, and the second and third patterns each involve only four species. The probabilities for these three patterns have already been given above. The probability of obtaining the fourth pattern, +++++, is approximately equal to

$$P_{5+} = a_{00}[p(2t_1 + t_2 + t_3 + t_4)^r p(t_2 + t_2 + t_4)^r p(t_2)^r + F(2t_1 + t_2 + t_3 + t_4, t_2, t_2 + t_3 + t_4)]p(t_3 + t_4)^r p(t_3)^r p(t_4)^{2r} + a_{00}p(2t_1 + t_2 + t_3 + t_4)^r p(t_2 + t_3 + t_4)^r \cdot [F(t_2, t_3, t_3 + t_4)p(t_4)^{2r} + p(t_2)^r p(t_3 + t_4)^r F(t_3, t_4, t_4)].$$

Table 8 shows the relative probabilities of the 20-site distribution patterns, given that the restriction site is present in two or three of the five species. Note that the relative probabilities are complicated functions of the λt_i values. However, for the λt_i values considered, either the first (+++−−) or the second pattern (++−−−) is the most frequent. Patterns 3, 4 and 5 (−−−++, −−+++ and ++−+−) usually occur with intermediate frequencies among the 20 patterns, but the relative frequency of pattern 4 may become low as λt_1 or λt_2 becomes relatively large. Patterns 13 to 20 generally occur with low frequencies.

With five species there are 15 possible networks, each of which is compatible with four of the 20-site distribution patterns. A network will occur with a high frequency if it is compatible with the most frequent patterns of site distributions. The correct network, (AB)-C-(DE), is compatible with the first four patterns in Table 8 and is expected to be the most frequent. The networks (AB)-D-(CE), (AC)-B-(DE), A-(BC)-(DE) and (AB)-(CD)-E are each compatible with two of the first four patterns and may occur with a high frequency. In addition, there are ten other possible networks. Therefore, the probability of obtaining the correct network may be low under many circumstances.

To evaluate the probability (P^+) of obtaining the correct network for a given number (N) of informative sites, a computer simulation can be conducted as follows. Let us use the case of $\lambda t_1 = \lambda t_2 = \lambda t_3 = \lambda t_4 = 0.010$ as an example. To simulate an informative site, a uniform pseudo-random number is generated to select one of the 20 possible site distribution patterns according to the probabilities given in the first row in Table 8. The simulation is repeated $M = 120{,}000$ times. To evaluate P^+ for a given N, we have $M/N = 6000$ replicates for $N = 20$, $M/N = 4000$ replicates for $N = 30$, and so on. In each replicate, we give the score 1 if network 1 is the most compatible network, but give 0 if any of the other 14 networks is more frequent than or as frequent as network 1. Then P^+ is equal to the total score divided by the number of replicates.

The simulation result is shown in Table 9. In the first case, $\lambda t_1 = \lambda t_2 = \lambda t_3 = \lambda t_4 = 0.010$, and P^+ is 0.94 or larger for $N \geq 20$. Therefore, if the λt_i values are more or less equal and of order 0.01, then it is quite certain that the inferred network is the true network, provided that there are more

TABLE 9

Probability (P^+) of obtaining the correct network and frequencies of the five most frequent networks for a given number (N) of informative sites

							Networks[c]					
λt_1	λt_2	λt_3	λt_4	P_I[a] ($\times 10^5$)	N	P^+[b]	1	2	3	4	5	Others
0.010	0.010	0.010	0.010	7	20	0.94	0.96	0.02	0.00	0.01	0.01	0.001
					30	0.98	0.99	0.01	0.00	0.00	0.01	0.000
0.010	0.010	0.010	0.050	14	20	0.29	0.37	0.12	0.10	0.10	0.08	0.227
					30	0.40	0.47	0.11	0.09	0.10	0.07	0.154
					40	0.49	0.56	0.10	0.08	0.09	0.06	0.110
					200	0.95	0.96	0.02	0.00	0.01	0.01	0.001
0.025	0.025	0.010	0.050	17	20	0.39	0.49	0.06	0.16	0.16	0.04	0.100
					30	0.51	0.59	0.04	0.14	0.15	0.02	0.052
					40	0.59	0.66	0.03	0.13	0.14	0.01	0.029
					200	0.95	0.96	0.00	0.03	0.02	0.00	0.000
0.025	0.025	0.010	0.025	14	20	0.67	0.75	0.02	0.10	0.11	0.01	0.012
					30	0.80	0.84	0.01	0.07	0.08	0.00	0.002
					50	0.91	0.93	0.00	0.03	0.04	0.00	0.000
0.025	0.025	0.025	0.025	16	20	0.78	0.84	0.07	0.02	0.02	0.05	0.005
					50	0.97	0.97	0.01	0.00	0.00	0.01	0.000
0.012	0.006	0.022	0.044	15	30	0.43	0.50	0.23	0.02	0.02	0.18	0.058
					40	0.50	0.56	0.22	0.01	0.01	0.17	0.032
					200	0.84	0.85	0.08	0.00	0.00	0.07	0.000
					400	0.95	0.96	0.03	0.00	0.00	0.01	0.000

It is assumed that $r = 6$ and that nucleotide substitution occurs randomly.

[a] P_I is the probability that a randomly chosen sequence of r nucleotides is an informative site.

[b] P^+ is the probability that network 1 is strictly more frequent than any of the other networks.

[c] The first five networks are 1, (AB)-C-(DE); 2, (AB)-D-(CE); 3, (AC)-B-(DE); 4, A-(BC)-(DE); and 5, (AB)-(CD)-E. When a tie for the most compatible network occurs among n networks in a replicate, $1/n$ of the replicate is assigned to each of the n networks. Network 1 is the correct network.

than 20 informative sites available. In the case of mammalian mtDNAs, to have 20 informative sites for $P_I = 7 \times 10^{-5}$ (Table 9), one requires about $20/(P_I \times 16{,}500) = 17$ six-base enzymes (see above for the computation). This is feasible. In case 2, λt_4 increases to 0.05, whereas the other λt_i's remain the same. In this case, P^+ is lower than 0.5 even when N is as large as 40, and for P^+ to be 95% or larger, one requires 200 informative sites or about 170 six-base enzymes. This is impracticable. In case 3, λt_1 and λt_2 increase to 0.025, whereas λt_3 and λt_4 remain the same as in case 2. The P^+ value is considerably larger than that in case 2 when N is small, but increases more slowly as N increases, so that for P^+ to be 95% or larger, one also requires 200 informative sites. In case 4, λt_4 is reduced to 0.025, so that the difference between λt_3 and λt_4 becomes smaller. The number of informative sites and the number of six-base enzymes required for P^+ to be 95% or larger are now reduced to 55 and 24, respectively. In case 5, $\lambda t_1 = \lambda t_2 = \lambda t_3 = \lambda t_4 = 0.025$, and the number of informative sites and the number of six-base enzymes required for P^+ to be 95% or larger are about 48 and 18, respectively. In this case the performance of the compatibility method is about as good as in case 1. In case 6, the λt_i values are similar to those in the phylogeny in Figure 3. The number of informative sites required for P^+ to be 95% or larger is extremely large. For $N = 30$, P^+ is only 0.43 (Table 9). In FERRIS, WILSON and BROWN's (1981) data, N is only 31 (Table 5). Therefore, this set of data does not seem to be large enough for drawing a definite conclusion about the phylogenetic relationships among the five primate species.

Table 9 also shows the frequencies of the five most frequent networks. The results were obtained as follows. In each sample of N informative sites, the score for a given network is 1 if it is more compatible with the informative sites than all other networks, $1/n$ if it ties with $n - 1$ other networks for the most compatible network, and 0 in all other cases. The frequency of a network is then equal to its total score divided by the number of replicates. Network 1 is the correct network, whereas networks 2, 3, 4 and 5 each have a clustering error; for example, in network 3 [(AC)-B-(DE)], A is grouped together with C rather than with B. In all the cases in Table 9, each of the four latter networks is considerably less frequent than network 1; however, in some cases, the sum of their frequencies is close to or larger than the frequency of network 1. Also, in case 2, the sum of the frequencies of the five most frequent networks is less than 90%, even if N is as large as 40. Nevertheless, except for this case, the sum of the frequencies of the first five networks is 95% or higher, even if N is only 30 or smaller. This means that, in most of the cases given in Table 9, the compatibility method will, with a high probability, give the correct network or a network with only one clustering error. From this point of view, it may be regarded as a useful method for inferring phylogenetic relationships from restriction site data. Moreover, even in a situation where the method does not have a high probability of identifying the correct network, it may still provide useful information about the plausibility of a particular network. For example, in case 6, network 3 occurs with probability 0.03 if $N = 20$ (not shown) and with probability 0.02 if $N = 30$ (Table 9). Thus, this network

(Figure 3c) does not seem plausible for the phylogenetic relationships among the five primate species. However, this computation does not warrant the rejection of network 3 because it involves a number of assumptions, one of which is that the λt_i values used are the true values. If there is any deviation from this assumption, the frequency of network 3 may become substantially different from those given above because it is highly dependent on the λt_i values.

We note from Table 9 that, for $N = 30$, P^+ is 0.40 for case 2 and 0.43 for case 6; yet for P^+ to be 95%, case 6 requires two times more informative sites than does case 2. This seemingly puzzling result can be explained as follows. In case 6, the frequencies of networks 2 and 5 for $N = 30$ are as high as 23% and 18%, respectively. Therefore, the occurrence of either network 2 or 5 cannot be neglected until N becomes extremely large. Indeed, these two networks occur with frequencies 8 and 7%, respectively, even when N is as large as 200. In case 2, although the frequencies of networks 2, 3, 4 and 5 for $N = 30$ are quite high, none of them is as high as those of networks 2 and 5 in case 6. Consequently, as N increases, the frequency of each of the four networks is expected to be reduced at a rate faster than those for networks 2 and 5 in case 6. In conclusion, P^+ will increase at a slow rate if any of networks 2–15 has a fairly high probability of occurrence for a small N. Such a situation can occur when λt_4 is relatively large and one or more of the other λt_i's are small. For example, in case 6, λt_4 is 0.044 but λt_2 is only 0.006, so that there is a good chance for C and D or C and E to be clustered together. As a consequence, networks 5 [(AB)-(CD)-E] and network 2 [(AB)-D-(CE)] occur with high frequencies when N is small.

It is clear from Table 9 that the performance of the unordered compatibility method is highly dependent on the λt_i values. As in the case of four species, it performs well when the λt_i values are more or less equal and relatively small. (Of course, the number of informative sites required for P^+ to be 95% increases with the number of species under study because the chance of making a clustering error increases as the number of species increases.) It is also expected to perform well if λt_1 and λt_4 are small and λt_2 and λt_3 are relatively large. It performs poorly when λt_4 is large and one or more of the other λt_i's are small.

DISCUSSION

In this study I have chosen to examine the unordered compatibility method for two reasons. First, it is simple and its statistical properties can be examined analytically. Second, the network inferred by the unordered compatibility method is often the same as the maximum parsimony network. For example, this is true for the case of the mtDNA data from the five primate species considered above. Actually, in the case of four species the two methods are identical. Thus, the above conclusion about the performance of the unordered compatibility method might hold approximately for the performance of maximum parsimony methods.

In the present study I have assumed that the rate of nucleotide substitution is the same for all regions of the DNA sequence and is constant over time. Further, the model used requires that the stationary frequencies of A, T, C and G all equal ¼. These assumptions are unrealistic, and a further study should be made without these assumptions. However, it is clear from the present results that the unordered compatibility method is unlikely to perform well if some of the branches between ancestral nodes are considerably shorter than the branches leading to more recent species.

As illustrated in this study, it is important to know the strengths and weaknesses of a method, because a method may perform well under one set of conditions but poorly under another. Through this type of study, we may find methods that are complementary to each other in their performance. Then, the proliferation of alternative methods for reconstructing phylogenies could become a boon rather than "a plague" (FELSENSTEIN 1984).

I thank J. FELSENSTEIN for detailed comments and suggestions and P. PAMILO, N. SAITOU, P. M. SHARP, J. C. STEPHENS and P. SMOUSE for discussion and suggestions. This work was supported by National Science Foundation grant BRS 8303965.

LITERATURE CITED

ANDERSON, S., A. T. BANKIER, B. G. BARRELL, M. H. L. DE BRUIJN, A. R. COULSON, J. DROUIN, I. C. EPERON, D. P. NIERLICH, B. A. ROE, F. SANGER, P. H. SCHREIER, A. J. H. SMITH, R. STADEN and I. G. YOUNG, 1981 Sequence and organization of the human mitochondrial genome. Nature **290:** 457–464.

AOKI, K., Y. TATENO and N. TAKAHATA, 1981 Estimating evolutionary distance from restriction maps of mitochondrial DNA with arbitrary G + C content. J. Mol. Evol. **18:** 1–8.

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.

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.

BROWN, W. M., M. GEORGE, JR. and A. C. WILSON, 1979 Rapid evolution of animal mitochondrial DNA. Proc. Natl. Acad. Sci. USA **76:** 1967–1971.

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.

CANN, R. L., W. M. BROWN and A. C. WILSON, 1984 Polymorphic sites and the mechanism of evolution in human mitochondrial DNA. Genetics **106:** 479–499.

DEBRY, R. W. and N. A. SLADE, 1985 Cladistic analysis of restriction endonuclease cleavage maps within a maximum-likelihood framework. Syst. Zool. **34:** 21–34.

EDWARDS, A. W. F. and L. L. CAVALLI-SFORZA, 1964 Reconstruction of evolutionary trees. pp. 67–76. In: *Phenetic and Phylogenetic Classification* (Syst. Assn. Publ. 6), Edited by V. H. HEYWOOD and J. MCNEILL. Systematics Association, London.

ENGELS, W. R., 1981 Estimating genetic divergence and genetic variability with restriction endonucleases. Proc. Natl. Acad. Sci. USA **78:** 6329–6333.

ESTABROOK, G. F. and F. R. MCMORRIS, 1980 When is one estimate of evolutionary relationships a refinement of another? J. Math. Biol. **10:** 367–373.

FELSENSTEIN, J., 1984 The statistical approach to inferring evolutionary trees and what it tells us about parsimony and compatibility. pp. 169–191. In: *Cladistics: Perspectives on the Reconstruction of Evolutionary History*, Edited by T. DUNCAN and T. F. STUESSY, Columbia University Press, New York.

FELSENSTEIN, J., 1985 Confidence limits on phylogenies with a molecular clock. Syst. Zool. **34:** 152–161.

FERRIS, S. D., R. D. RAGE, E. M. PRAGER, U. RITTE and A. C. WILSON, 1983 Mitochondrial DNA evolution in mice. Genetics **105:** 681–721.

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

FITCH, W. M., 1967 Evidence suggesting a non-random character to nucleotide replacements in naturally occurring mutations. J. Mol. Biol. **26:** 499–507.

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

GOJOBORI, T., W.-H. LI and D. GRAUR, 1982 Patterns of nucleotide substitution in pseudogenes and functional genes. J. Mol. Evol. **18:** 360–369.

KAPLAN, N. and C. H. LANGLEY, 1979 A new estimate of sequence divergence of mitochondrial DNA using restriction endonuclease mapping. J. Mol. Evol. **13:** 295–304.

KIMURA, M., 1980 A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. J. Mol. Evol. **16:** 111–120.

LE QUESNE, W. J., 1969 A method of selection of characters in numerical taxonomy. Syst. Zool. **18:** 201–205.

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

LI, W. -H., C. -I. WU and C. -C. LUO, 1984 Nonrandomness of point mutation as reflected in nucleotide substitutions in pseudogenes and its evolutionary implications. J. Mol. Evol. **21:** 58–71.

LI, W. -H., C. -I. WU and C. -C. LUO, 1985 A new method for estimating synonymous and nonsynonymous rates of nucleotide substitution considering the relative likelihood of nucleotide and codon changes. Mol. Biol. Evol. **2:** 150–174.

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. and F. TAJIMA, 1985 Evolutionary change of restriction cleavage sites and phylogenetic inference for man and apes. Mol. Biol. Evol. **2:** 189–205.

POWELL, J. R., 1983 Interspecific cytoplasmic gene flow in the absence of nuclear gene flow: evidence from *Drosophila*. Proc. Natl. Acad. Sci. USA **80:** 492–495.

SNEDECOR, G. W. and W. G. COCHRAN, 1967 *Statistical Methods*, Ed. 6. Iowa State University Press, Ames.

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., 1983a Convergent evolution and nonparametric inferences from restriction data and DNA sequences. pp. 151–179. In: *Statistical Analysis of DNA Sequence Data,* Edited by B. S. WEIR. MARCEL DEKKER, New York.

TEMPLETON, A. R., 1983b 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., 1986 Relation of humans to African apes: a statistical appraisal of diverse types of data. In: *Evolutionary processes and Theory,* Edited by E. NEVO and S. KARLIN. Academic Press, New York. In press.

UPHOLT, W. B., 1977 Estimation of DNA sequence divergence from comparison of restriction endonuclease digests. Nucleic Acids Res. **4:** 1257–1265.

Communicating editor: B. S. WEIR