J Mol Evol (1987) 26:123–131

Journal of Molecular Evolution

Estimation of Hominoid Phylogeny from a DNA Hybridization Data Set

Joseph Felsenstein

Department of Genetics SK-50, University of Washington, Seattle, Washington 98195, USA

Summary. Analysis of the expanded data set of Sibley and Ahlquist (1987) on primate phylogeny using a maximum likelihood mixed model analysis of variance method shows that there is significant evidence for resolving the *Homo–Pan–Gorilla* trifurcation in favor of a *Homo–Pan* clade. The resulting tree is close to that estimated by Sibley and Ahlquist (1984). The mixed model can be used to test a number of hypotheses about the existence of components of variance and the linearity of the relationship between branch length and expected distance. No evidence is found that there is a variance component for extract, or for the individual from which the extract was taken. A variance component for experiment does seem to exist, presumably arising as a result of error of measurement of the common standard from which all values in the same experiment were subtracted. There is significant evidence that the relationship between total branch length between species and their expected distances is nonlinear, or else that the measurement error on larger distances is greater than on smaller ones. Allowing for the nonlinearity might cause one to infer the time of distant common ancestors as less remote than the measured hybridization values would imply if used directly.

Key words: DNA hybridization – Phylogeny – Hominoids – Statistical analysis – Maximum likelihood – Mixed model ANOVA

Introduction

Sibley and Ahlquist (1984) used DNA hybridization of a single-copy fraction of nuclear DNA to examine the phylogeny of hominoids. Their results have attracted wide attention and controversy because they concluded that, within the *Pan–Homo–Gorilla* clade, humans and chimpanzees were a monophyletic group. Although in other respects their phylogeny was consistent with morphological evidence, the intrinsic importance of this part of the phylogeny (at least from our perspective as humans) and the new perspective it would cast on such issues as the evolution of bipedal locomotion, make its reexamination essential. Brown et al. (1982) and Templeton (1983, 1985) have argued that data from mitochondrial restriction sites and nucleic acid sequences resolves the trichotomy differently, *Pan* and *Gorilla* forming a clade, which would allow for the unique and unreversed derivation of knuckle-walking.

In the accompanying paper, Sibley and Ahlquist (1987) have presented an expanded data set, more than doubling the number of DNA hybridizations. This paper presents a statistical analysis of those data, taking into account the major sources of correlation between observations and examining some of the objections that have been raised to previous analyses.

Structure of the Data

The data are given in Table 1 of Sibley and Ahlquist (1988). The details of the experimental methods are given by those authors. There were 514 data points in all, of which 450 were used in the present analysis. Each represents a $\Delta T_{50}H$ value between a tracer DNA and a driver DNA. The $\Delta T_{50}H$ value is the difference in the temperatures at which half of the tracer–driver hybrid melts and the temperature at which half of a standard melts, with the standard consisting of the same sample used as both tracer and driver. It would

Offprint requests to: J. Felsenstein

have been preferable to have recorded the actual melting temperatures of both the experimental hybrids and their standards; this would have enabled us to tell to what extent correlations among measurements from the same experiment could be attributed to error in the measurement of the standard. The data analyzed here are the $\Delta T_{50}H$ values.

For each data point, the species and the individual from which the tracer was extracted, the number of the extract (some individuals were sampled more than once), and the same pieces of information for the driver DNA are known. The number of the experiment was also recorded. This is important because an experiment represents 25 or fewer hybrids whose $\Delta T_{50}H$ values were all calculated from a common standard. Experimental error in measuring the standard would be expected to cause correlated errors in the $\Delta T_{50}H$ values from each experiment.

The 514 data points were reduced to 450 for this analysis by omitting 64 data points, which represent DNA hybrids between hominoids and a variety of cercopithecoids. To avoid complications caused by heterogeneity of the outgroup species, only one cercopithecoid, the hamadryas baboon (*Papio hamadryas*), was retained, because it had the most data points of any cercopithecoid and because it was the only one from which a tracer DNA was extracted.

Statistical Model

Each $\Delta T_{50}H$ value, henceforth called a distance, can be considered to be the sum of an expectation and an error around that expectation. The distribution of distances will be assumed to be multivariate normal. Each distance has an expectation which is assumed to be a function of the sum of branch lengths between those two species in the unknown true phylogeny. Around this expectation there is a statistical error. This represents individual measurement error, plus components of error common to distances sharing the same tracer DNA extract, the same driver DNA extract, and standardized against the same standard. If distance D_{ijklmn} is measured between extract k of species i and extract l of species j, standardized against standard m, and represents replicate measurement n of that particular combination, then

$$Y(D_{ijklmn}) = d_{ij} + \beta_{ik} + \gamma_{jl} + \delta_m + \epsilon_{ijklmn} \quad (1)$$

where

- Y is a transformation reflecting the nonlinearity of the dependence of distance on total intervening branch length, intended to return the observed distance D to a scale on which branch lengths are additive,
- d_{ij} is the sum of the branch lengths in the phylogeny between species i and species j,
- β_{ik} is the error common to all measurements made with tracer DNA k extracted from species i,
- γ_{jl} is the error common to all measurements made with driver DNA l extracted from species j,
- δ_m is the error common to all measurements which are relative to the same experimental standard, and hence are assigned the same experiment number,
- ϵ_{ijklmn} is the individual measurement error not attributable to any of these causes.

The errors β, γ, δ, and ϵ have zero expectation. The expectation of $Y(D_{ijklmn})$ is d_{ij}, which is the sum of branch lengths between species i and species j. The statistical model is a mixed model analysis of variance, with the fixed effects being the branch lengths and the random effects being the β, γ, δ, and ϵ. The $Y(D_{ijklmn})$ are multivariate normally distributed, and the Ds themselves have a density function which is calculable from knowledge of the transformation Y.

My interest is in calculating the fixed effects for various possible tree topologies, in estimating the variance components corresponding to the errors β, γ, δ, and ϵ, and in computing likelihoods to test various hypotheses concerning the transformation, the phylogeny, and the magnitudes of the variance components. I am, of course, particularly interested in the details of the phylogeny in the *Homo–Pan–Gorilla* region.

The transformation Y(D) reflects the nonlinearity of the relationship between the expected number of changes in the DNA and the observed distance. I use a model like that of Jukes and Cantor (1969) which predicts nonlinearity between the number of changes and the fraction of sites at which the DNAs differ, owing to multiple "hits" overlaying each other at one site. The corresponding transformation is the inverse of

$$D = (1 - e^{-aY})/a \quad (2)$$

so that D is initially a linear function of branch length, but then approaches as asymptotic value of 1/a as Y becomes large. Equation (2) defines Y as:

$$Y(D) = -\ln(1 - a\,D)/a. \quad (3)$$

For the purposes of this paper, the important feature of Eq. (2), seen when it is expanded as a Taylor series, is that it is approximately a quadratic function of Y:

$$D \cong Y - aY^2/2. \quad (4)$$

The values of Y encountered in this data set are small, and thus terms beyond the quadratic contribute little. Any other transformation which included a similar quadratic term would thus probably

do as well. The basic point of the transformation is simply to allow a curvilinear dependence of D on branch length. I would have used Eq. (4) directly, but it was less tractable mathematically than Eq. (2).

The log likelihood for the Ds is of the form:

$$\ln L = -\tfrac{1}{2} \ln|\mathbf{V}| - \tfrac{1}{2}(\mathbf{Y} - \mathbf{d})'\mathbf{V}^{-1}(\mathbf{Y} - \mathbf{d}) - a \sum_{ij} Y_{ij} \quad (5)$$

where the final term is the Jacobian correction for the transformation from D to Y. The elements of the expectation vector **d** are the sum of branch lengths, and the covariance matrix **V** is generated by the model [Eq. (1)]. The statistical analysis of the Ds involves maximizing the log likelihood [Eq. (5)] by fitting branch lengths (for a given tree topology), variance components, and the nonlinearity parameter a. Different hypotheses about these parameters can be tested by performing likelihood ratio tests.

The variance components corresponding to the four terms on the right-hand side of Eq. (1) are σ_I^2, σ_J^2, σ_X^2, and σ_E^2, these being the tracer extract, driver extract, experiment, and error components.

Computational Methods

When all four variance components are included, it is not a simple matter to maximize the log likelihood [Eq. (5)]. I have used a variant of an EM-algorithm (Dempster et al. 1977, pp. 17–18) implemented so as to iteratively improve one variance component after another, recalculating the maximum likelihood estimates of the branch lengths after each variance component has gone through one cycle of the EM-algorithm. The details of the algorithm will not be given here.

When the variance components σ_X^2 and σ_E^2 are the only ones present, the analysis is considerably simpler. In that case the covariance matrix **V** is block diagonal, and each block has a simple structure that enables $\mathbf{V}^{-1}$ to be computed analytically. The model is in effect an unbalanced one-way analysis of variance with a particular linear model for the means.

Let **v** be the vector of branch lengths in the phylogeny. Suppose that a design matrix **T**, dependent on the tree topology, is given so that the expected transformed distances d are

$$\mathbf{d} = \mathbf{T}\mathbf{v}. \quad (6)$$

Equation (5) is then of the form

$$\ln L = -\tfrac{1}{2} \ln|\mathbf{V}| - \tfrac{1}{2}(\mathbf{Y} - \mathbf{T}\mathbf{v})'\mathbf{V}^{-1}(\mathbf{Y} - \mathbf{T}\mathbf{v}) - a \sum_{ij} Y_{ij} \quad (7)$$

and when this is maximized with respect to the elements of **v** the normal equations are:

$$\mathbf{T}'\mathbf{V}^{-1}\mathbf{T}\mathbf{v} = \mathbf{T}'\mathbf{V}^{-1}\mathbf{Y}. \quad (8)$$

In the present case this is a set of at most 13 simultaneous linear equations in 13 unknowns, and this is not difficult to solve numerically. Thus, for given values of the variance components contributing to **V**, the likelihood for the branch lengths v_i could be maximized.

The maximization of the likelihood for the variance components is less straightforward. If the ith experiment has n_i measurements in it, the jth of these being called Y_{ij} and its expectation being d_{ij}, it turns out that the likelihood [Eqs. (5) and (7)] can be rewritten as

$$\ln L = -\tfrac{1}{2} \sum_i [(n_i - 1)\ln \sigma_E^2 + \ln(\sigma_E^2 + n_i\sigma_X^2)] - \tfrac{1}{2} \sum_i \left[\sum_j (Y_{ij} - d_{ij})^2/\sigma_E^2 - \frac{\sigma_X^2(\sum [Y_{ij} - d_{ij}])^2}{\sigma_E^2(\sigma_E^2 + n_i\sigma_X^2)} \right] \quad (9)$$

which, although it cannot be analytically maximized with respect to σ_E^2 and σ_X^2, can easily be maximized numerically.

The strategy has been iterative. Starting with initial estimates of the variance components σ_X^2 and σ_E^2, I solved Eq. (8) to obtain an estimate of the v_i. These then were used to generate the d_{ij} in Eq. (9), which was maximized to obtain new estimates of the variance components, and so on, iteratively, until the process converged. This brings one to a stationary point on the likelihood surface, which in practice is always the maximum. The parameter *a* of the transformation has been iterated by direct search, with a resolution of 0.01. For each value of *a* the other parameters were iterated to maximize the likelihood.

Results

Table 1 shows a summary of the average distances between the species; a tree can be roughly estimated from this by visual inspection. In fact, the maximum likelihood phylogeny has that same tree topology. Figure 1 shows the maximum likelihood tree, the branch lengths being the horizontal dimension. The tree has been rooted by requiring *Papio* to be at the same height as *Pan troglodytes*; this rooting is also consistent with the use of *Papio* as an outgroup. The parameter values, other than the branch lengths, and the log likelihood are shown in the first line of Table

Table 1. Average DNA hybridization distances among the eight species in this data set

	Hs	Pt	Pp	Gg	Po	Hy	Hl	Ph
Hs	–							
Pt	1.628	–						
Pp	1.645	0.689	–					
Gg	2.267	2.210	2.367	–				
Po	3.600	3.576	3.562	3.550	–			
Hy	4.700	5.133	4.200	4.543	4.933	–		
Hl	4.779	4.760	5.000	4.753	4.745	1.950	–	
Ph	7.330	7.336	6.967	7.078	7.486	–	7.100	–

The species symbols are Hs = *Homo sapiens*, Pt = *Pan troglodytes*, Pp = *Pan paniscus*, Gg = *Gorilla gorilla*, Po = *Pongo pygmaeus*, Hl = *Hylobates lar*, Hy = *Hylobates syndactylus*, and Ph = *Papio hamadryas*. The entries are the averages of all data points for each species pair from Table 1 of Sibley and Ahlquist (1987)

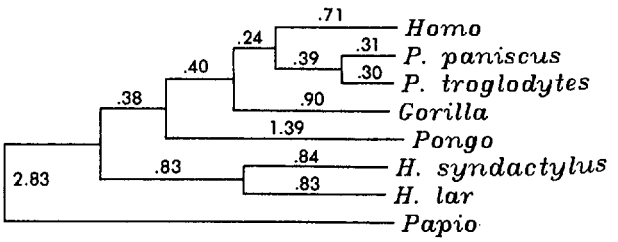

Fig. 1. The maximum likelihood estimate of the hominoid phylogeny under the mixed model analysis, case 1 of Table 2. The abbreviation "H." stands for *Hylobates*, and "P." for *Pan*.

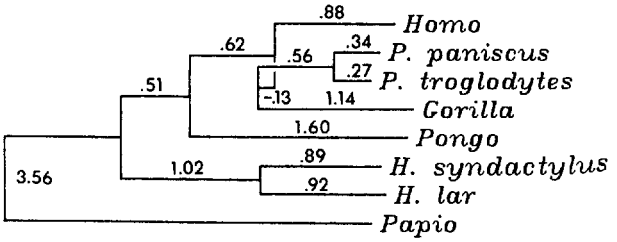

Fig. 2. The same case as Fig. 1 but with the *Pan–Gorilla* clade forced to exist, and negative branch lengths allowed (case 2 of Table 2)

2. The tree is of the same topology as the original tree of Sibley and Ahlquist (1984) and contains the *Pan–Homo* clade.

Cases 2 and 3 of Table 2 show two alternative topologies. Figures 2 and 3 show these trees. Case 2 has a negative branch length. Case 3 has a trifurcation, the result of constraining the branch lengths to be nonnegative within the second topology. A similar constraint on the first topology has no effect, as the maximum likelihood tree has all positive branch lengths. Figure 4 shows the tree obtained from case 15 with a molecular clock and transformation of the distances, but without extract effects.

Testing Tree Topology

It is possible to use likelihood ratio tests to test tree topologies, although as I have explained elsewhere (Felsenstein 1983) there are statistical complications. The trifurcation (tree III) can be tested against the maximum likelihood tree, as this difference amounts to constraining one parameter (the length of the branch leading to the *Pan–Homo* clade) to be zero. The test thus has 1 degree of freedom, and involves doubling the difference between the log likelihood values of the trees and comparing this to the percentiles of a chi-square distribution with 1 degree of freedom. The chi-square variate is thus $2 \times (360.03 - 296.54) = 126.98$. This is too large to locate on a standard chi-square table. The simplest way to compute its significance is to note that since a chi-square variate with 1 degree of freedom is the square of a standard normal variate, the significance of this should be the same as that of a normal variate 11.27 standard deviations away from its expectation, tested with a two-tailed test. This gives $p = 1.92 \times 10^{-31}$, a strongly significant value.

Testing tree I against tree II is not directly possible, since these are not nested hypotheses (neither is a subcase of the other). I have discussed elsewhere in the analogous case of a least-squares test (Felsenstein 1984, 1986) the difficulties in this test. A conservative test can be made by substituting the log likelihood for tree II for that for tree III and assuming 1 degree of freedom. This would be expected to reduce the significance of the result. In the present case it results in a chi-square value of $2 \times (360.03 - 308.04) = 103.98$, a normal deviate of 10.20, and $p = 1.94 \times 10^{-26}$.

The remainder of the lines in Table 2 show cases in which there was no transformation of the data, or various variance components were set to zero, or a molecular clock was assumed. Note that whichever of these is assumed, the difference between the likelihoods of tree I and either tree II or III are always highly significant. The comparison least favorable to tree I is that between cases 11 and 12, which gives a chi-square value of 82.74, a normal deviate of 9.10, and $p = 9.36 \times 10^{-21}$.

Testing the Transformation of Distances

A nonlinear transformation [Eqs. (3) and (4) above] has been used to relate distances to branch lengths. Whether the transformation is necessary can be tested by comparing log likelihoods with and without the transformation. This restricts the value of one parameter to zero, and thus also has 1 degree of freedom. The relevant comparison is of cases 1 and 8. This gives a chi-square value of 32.58, a normal deviate of 5.71, and $p = 1.1 \times 10^{-8}$. The nonlinearity of the distances is thus strongly sup-

Table 2. The maximum log likelihoods achieved under various models and tree topologies

Case	Clock?	Tree	a	σ_I^2	σ_J^2	σ_X^2	σ_E^2	ln L
1	No	I	−0.19	0.0000	0.0005	0.0054	0.0256	360.03
2	No	II	−0.09	0.0000	0.0001	0.0127	0.0501	308.04
3	No	III	−0.07	0.0000	0.0002	0.0152	0.0579	296.54
4	No	I	−0.19	0	0	0.0049	0.0265	359.52
5	No	II	−0.09	0	0	0.0120	0.0506	308.06
6	No	III	−0.07	0	0	0.0145	0.0587	296.48
7	No	I	−0.18	0	0	0	0.0317	348.77
8	No	I	0	0.0000	0.0008	0.0138	0.0712	343.74
9	No	II	0	0.0000	0.0002	0.0206	0.0858	301.97
10	No	III	0	0.0000	0.0004	0.0227	0.0889	292.36
11	No	I	0	0	0	0.0130	0.0731	343.35
12	No	II	0	0	0	0.0197	0.0866	301.98
13	No	III	0	0	0	0.0218	0.0901	292.28
14	No	I	0	0	0	0	0.0838	332.86
15	Yes	I	−0.20	0	0	0.0048	0.0258	357.74
16	Yes	II	−0.03	0	0	0.0146	0.0836	271.11
17	Yes	III	−0.01	0	0	0.0174	0.1030	252.12
18	Yes	I	0	0	0	0.0137	0.0741	339.75
19	Yes	II	0	0	0	0.0174	0.1015	269.91
20	Yes	III	0	0	0	0.0185	0.1104	251.97

The tree topologies I, II, and III are those shown in Figs. 1–3. The lines of the table are numbered for ease of discussion in the text. Parameters given as 0 (rather than 0.0000) were held at 0 and not iterated

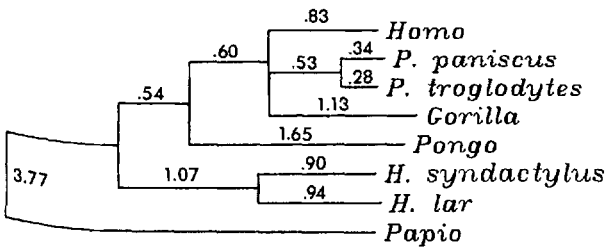

Fig. 3. The same case as Fig. 2, but with negative branch lengths not allowed (case 3 of Table 2). The branch leading to the *Pan–Gorilla* clade shrinks to length 0, so that there is, in effect, a trifurcation.

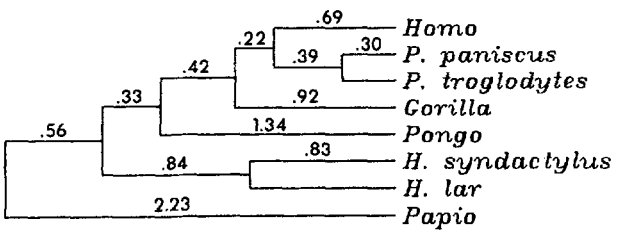

Fig. 4. The maximum likelihood estimate under the mixed model analysis with a molecular clock assumed and extract effects absent (case 15 of Table 2)

ported. Similar tests carried out under the other tree topologies are also significant (cases 2, 3, 5, and 6 versus cases 9, 10, 12, and 13, respectively) unless a molecular clock is assumed (cases 16 and 17 versus lines 19 and 20).

The direction of the transformation is surprising. Although Eq. (2) is derived from considerations of superposition of random changes in DNA, that assumes that the constant a is positive. Its maximum likelihood estimate is negative, so that the observed distance curves upwards with increasing branch length. Note that the effect of this transformation is to make the remote forks less remote than might be imagined from their observed distances. In the tree of Fig. 4 the *Homo–Pan* fork is 1.388 units of time in the past, while the *Homo–Hylobates* split is 3.337 units of time ago, 2.4 times as distant. Without any transformation these numbers would have been 1.598 and 4.766, for a ratio of 2.98. Thus if time is calibrated by the smaller distances, the effect of the transformation is to make the remote forks less remote; if it is calibrated by the remote forks, its effect is to make the recent forks less recent. The matter seems worth attention given the difficulties of reconciling dates of fossils with some suggested calibrations of the molecular clock. However, it is also possible that the transformation reflects a greater measurement error on the larger values of $\Delta T_{50}H$. This would be confounded with the transformation of branch length, and there is no way to separate these two effects given the present data.

Testing for Extract Effects

The statistical model allows there to be an effect of the individual extract of DNA on the measured distance. This can be tested by restricting either the variance component σ_I^2 or σ_J^2 to be zero and seeing what effect this has on the likelihood. In fact, the maximum likelihood estimates of σ_I^2 always turned out to be zero, so that there is no evidence that this variance component is nonzero. The estimates for σ_J^2 are quite small. A likelihood ratio test can be

done with 1 degree of freedom by comparing cases 1 and 4 (see also cases 8 and 11). The chi-square value is 1.02, which gives $p = 0.31$, a nonsignificant value. There seems to be no evidence that it matters which extract is used for a given species. Since most of the extracts are from different individuals, the individual effect is expected to be nearly confounded with the extract effect, and so there is no evidence for an individual effect. Consideration of other topologies does not change this conclusion (compare cases 2, 3, 6, and 7 with cases 9, 10, 12, and 13, respectively).

Testing for Experiment Effects

The remaining shared component of variance in the mixed model is that due to the experiment, which is largely expected to be the result of all the values in one experiment being expressed as differences from a common standard. Measurement error in the standard should cause a correlation between values from the same experiment. The presence of this component of variance can be tested for by constraining the variance component σ_X^2 to be zero and seeing whether that results in a significant decrease in the log likelihood. This too has 1 degree of freedom. Comparing cases 4 and 7, a chi-square of 21.50 is obtained so that $p = 3.8 \times 10^{-6}$. Cases 11 and 14 show that a similar result is obtained if the distances are not transformed.

If the experimental measurements and their standards are equally variable, one can show that the experiment effects should account for half of the total variance. In the present analysis they account for only 17% of the variance. This seems to reflect to some extent a lower variability of the standards, which measure an individual extract's hybridization with itself, but primarily it reflects the fact that a biased estimate of the variance component σ_X^2 has been made. The present analysis is maximum likelihood (ML) rather than reduced (or restricted) maximum likelihood (REML). The former is expected to make unbiased estimates of the branch lengths but biased estimates of the first three variance components, while the latter makes unbiased estimates of all parameters. A crude estimate of the amount of bias suggests that an unbiased estimate of the experiment variance component would contribute about 33% of the total variance which is still not enough but at least in much closer agreement with the expectation.

Testing the Molecular Clock

In cases 15–20 in Table 2 the branch lengths of the estimated tree are constrained so that the tips are all at the same height (i.e., so that the expected distances from the root of the tree to the tips are all equal). This amounts to the assumption of a "molecular clock." When the branch lengths are constrained in this way, instead of 13 independent branch lengths there are only 6. This results from the constraint that certain branches be of equal length, and certain sums of branch lengths be equal. This gives one the opportunity to carry out a likelihood ratio test with 7 degrees of freedom.

The test compares cases 4 and 15, yielding a chi-square value of 3.56. Since there are 7 degrees of freedom in the test, $p = 0.83$ which is not significant. The comparable test when the distances are not transformed compares cases 8 and 18 and is also nonsignificant. However, when the other tree topologies are considered, these significantly reject the molecular clock (cases 5 and 6 compared with cases 16 and 17, respectively) and the same pattern is true when the distances are not transformed (cases 11, 12, and 13 compared with cases 18, 19, and 20).

Discussion

Inadequacies of This Analysis

There were a number of improvements possible in the analysis which I have not done because they would be computationally difficult and also seemed to promise little improvement over the present model. These include:

Absence of Individual Effects

The model [Eq. (1)] has effects for experiment and for the tracer and driver extracts, but not for the individual from whom the extract was taken. The extract effects thus might reflect either effects of the extract or effects of the individuals. There is some, but not much, information in the data to separate these effects: the same individual was often used for tracer and for driver extracts, and in a few cases several extracts were taken from one individual (there are three individuals—one human, one chimpanzee, and one gorilla that had more than one extract prepared). The effect for extracts and that for individuals was thus nearly confounded. The extract effects had no significant effect in any of the analyses, so that one infers that individual effects too are not provably present.

Handling of Transformation of Distances

The analysis using the transformations [Eqs. (2) and (3)] has assumed that the statistical errors act additively on a scale that is then transformed to yield the observed distances. Since the major components of error appear to be measurement errors, these might act additively on the scale of the ob-

served distances instead. I have discussed (Felsenstein 1986) the distinction between these two types of treatment; the particular method used was chosen for numerical tractability. However, since the estimates of the parameter a of the transformation in every case show the transformation to be only modestly (if significantly) nonlinear, it is unlikely that it matters whether the error is introduced before or after the transformation.

Asymptotic Nature of the Tests

Likelihood ratio tests are by their very nature asymptotic—only with sufficiently large amounts of data do the test statistics have the posited chi-square distribution. Here the number of data points is large enough (450) that it seems likely that asymptotic theory applies well. The closest case that could be checked exactly would be that of a one-way analysis of variance with 450 observations divided equally among 25 groups. The F distribution with 425 and 24 degrees of freedom has a 95% point of $F = 1.53$. The likelihood ratio test for the same hypothesis, that there is no group variance component, can be expressed in terms of F and is significant when $F = 1.79$, so that the likelihood ratio test is conservative. The strongly significant values obtained in the present analysis are thus unlikely to be artifacts of the asymptotic nature of the test.

Assumption of Normality of the Distances

The present model posits the distances to be drawn from a multivariate normal distribution. This is necessarily an approximation. Their true distribution is unknown. The extreme p values calculated above reflect the rapidity with which the tail of a normal distribution dies away as we get, say, 11.3 standard deviations out. It might be true that the distribution dies away more slowly. However, the Chebyshev inequality guarantees that any distribution with a finite expectation and finite variance has no more than $1/(11.3)^2 = 0.0078$ of its area beyond 11.3 standard deviations out. Thus, there is some confidence that the inappropriate use of a normal distribution is not entirely responsible for the significance of the results.

Failure to Record Standards

One of the limitations of this data set is that, as it was provided to me, the standards for each experiment had already been subtracted from the experimental values, and were not provided separately. If they were available, there would have been 25 more data points. This would have permitted an analysis of whether the experiment effect was entirely due to the measurement error of the standards, and whether the standards did in fact covary with the experimental points. This is, however, not a major concern of the present paper, and the amount of data added would have been small, so that this shortcoming of the data recording does not seem to have been serious.

Correlations between Distances

The most serious shortcoming of the present analysis is that it assumes that the only sources of error shared among the measured distances were experiment and extract effects. However, it is possible that random events in evolution were also shared and induced other correlations. For example, in the tree of Fig. 1, if we pass from *Papio* to *Homo*, and also from *Pongo* to *Gorilla*, these two paths share one segment in common (the branch leading to the ancestor of the *Homo–Pan–Gorilla* clade). If there is a burst of evolutionary change in this branch, that will increase all of the distances between the members of this clade and the other species. These are correlated variations in the distances, and my statistical model does not allow for this sort of variation.

There are two reasons why this will not cause trouble in the present analysis. First, with this kind of data one cannot distinguish between a random burst of evolution in an interior branch of the tree and a lengthening of that branch, unless a molecular clock is assumed. For most of the interior branches of the tree, their lengthening causes a departure from the clock. A burst of evolution can then be detected by the way it makes the tips of the tree fail to level. Without the clock assumption, such a burst of evolution cannot be detected. If these were sequence data rather than hybridizations, additional sequences might show whether the branch was long by giving an independent estimate of its length. With DNA hybridization, the length of the branch reflects the average number of changes in that branch in the relevant sequences, and more of the single-copy genome cannot be examined. In any case, it appears that there is no detectable departure from a molecular clock, which suggests with some confidence that the branch lengths are not much affected by randomness of base substitution.

The length of sequence that contributes to the estimate of branch length in any branch is large, and this leads one to expect that the randomness of base substitution will make a very small contribution to the statistical error of the observed distances. This is the second reason for confidence in the present model. A model calculation will serve to illustrate the point. Suppose that a hybridization value reflects the average number of base changes in a certain number of sequences, each of length 200 bases. With an average probability of base change of 5% (roughly in the range of the values seen in this analysis) each sequence of 200 bases will differ by an average of

10 bases. Using a binomial distribution, the variance of the number of differences is calculated to be 9.5, giving a standard deviation of 3.08 bases. Now assume that the experimental error is equivalent to a standard deviation of 10% of the measured value (again, roughly in the range of the present values). This would be equivalent to a standard deviation of one base.

In this numerical example, if measured distances reflect only one of these 200 base sequences, the variance of the observed distance would be 10.5, of which 9.5 would come from the randomness of base substitution and the rest from measurement errors. But if the distance represented the average of 10 sequences of length 200 bases each, then the two variance components would be 0.95 and 1.0; with 100 sequences, they would be 0.095 and 1.0; with 1000, 0.0095 and 1.0. If the hybridization values reflected sequence variation in as few as 100 sequences (20 kb), then the statistical variation in the distances is dominated by the experimental measurement error and is little affected by the randomness of base substitution, which has been averaged out. It is not hard to show that the same is true of the covariances and correlations between the distances.

The most serious potential weakness of the present data and analysis is that they might be measuring, not the whole of the single-copy DNA, but predominantly a few highly repeated sequences which have not been removed by the experimental procedures designed to do so. The above calculation suggests that even if the data reflect as few as 20 kb worth of independently evolving sequence, one still has enough to make the present statistical model appropriate. This suggests that the present analysis is not particularly vulnerable to the effect of multiple-copy sequences on the hybridization values.

Objections to Previous Analyses

There have been two major critiques of the statistical methods used in the original paper by Sibley and Ahlquist (1984), these being the papers of Templeton (1985) and Farris (1985). Templeton's paper has been rebutted by others (Ruvolo and Smith 1986; Saitou 1986) and defended by Templeton (1986). Templeton (1985) first reduced the data to rank orders of the pairwise mean distances between species and then applied nonparametric tests to these ranks. This was intended to gain robustness at the expense of power. Templeton's most recent (1986) reanalysis of these ranks argues that they lack power to discriminate between the major alternative topologies.

Even if his conclusion were accepted, it would not invalidate either the present paper or the analysis by Sibley and Ahlquist (1984), as it is evident that much of the information in the data is not retained when the data are reduced to rank-orderings of mean pairwise species distances. The significant differences found in the present data reflect both the inclusion of numerical magnitudes (rather than just ranks) and the details of the experimental design, including the replication within species and the organization of the points into experiments. The replication and error structure was not taken into account in Templeton's analysis. There thus seems little reason to regard Templeton's analysis as a serious objection.

The criticisms by Farris (1985) are a continuation of a critique he has had of all existing distance matrix methods for inferring phylogenies (Farris 1981). I have rebutted these (Felsenstein 1984; see also Farris 1986 and Felsenstein 1986). Farris (1985) found that the 1984 Sibley and Ahlquist data, analyzed by minimizing Fitch and Margoliash's (1967) Average Percent Standard Deviation (APSD) measure, support a tree having a *Pan–Gorilla* clade with a negative branch length leading to this clade. I have argued (Felsenstein 1986) that allowing negative branch lengths is inappropriate. It is best to constrain the estimated trees so as to avoid negative branch lengths, in which case the tree minimizing APSD is essentially the one found by Sibley and Ahlquist. Observation of a negative branch length is not, in itself, reason to reject the underlying statistical model. One must consider whether the negative branch length is significantly negative or whether it could have resulted from statistical measurement error. This has not been done by Farris (1985).

In the present data there are no negative branch lengths found, whether or not one assumes that the distances are nonlinear functions of branch length, unless the tree topology is forced to have the *Pan–Gorilla* clade and allowed negative branch lengths. In that case the tree found has considerably less than the maximum likelihood. Both the maximum likelihood tree and the tree minimizing the APSD measure (not shown here but nearly identical to Fig. 1) have the topology shown in Fig. 1. The present likelihood analysis, which takes into account different sources of correlation among distances, seems preferable to the least squares analysis, and like it, results in a tree without any negative branch lengths even when those are allowed.

Farris (1986) and I (Felsenstein 1984, 1986) are now in agreement on at least one point—that a major assumption of least-squares (and maximum likelihood) analyses of distance data is that the branch lengths add up linearly to give the expected distances between species. The present analysis not only incorporates a transformation to linearize the expected distances, but estimates its parameter (a), allow-

ing one to estimate and test the degree of nonlinearity. The results show that the relationship between branch length and expected distance may indeed be nonlinear, and in a direction different from what would be expected from Jukes and Cantor's (1969) formula.

Relative Power of Sequence and Distance Data

One may wonder how powerful DNA hybridization methods are, compared to direct sequencing methods. Although it is sometimes implied that DNA hybridization values, being based on very large numbers of sites, are far more accurate than sequencing methods, this does not take into account the measurement error of the hybridization values. A simple calculation is instructive. Accept Sibley and Ahlquist's (1984) calibration of 1 degree $\Delta T_{50}H$ as representing 1% nucleic acid sequence divergence. Then use this calibration to calculate, by summing the intervening branch lengths in Fig. 1, that *Pan troglodytes* and *Homo sapiens* differ at approximately 1.4% of their base positions. A single hybridization value has a standard error (estimated from line 1 of Table 2) of approximately 0.175, so that its coefficient of variation is 0.1255.

How much DNA must be sequenced to achieve the same coefficient of variation? If the DNA differs at a fraction p of its base positions, then on sequencing n bases, an average of np differences, with a (binomial) variance np(1 − p), will be seen. The coefficient of variation of the number of differences will be

$$C = \frac{[np(1-p)]^{1/2}}{np} = \frac{(1-p)^{1/2}}{n^{1/2}p^{1/2}}. \qquad (10)$$

Letting p = 0.014 and equating C to the hybridization value of 0.1255, then one can solve for n, the amount of sequence that would have equivalent statistical power. It turns out to be 4472 bases. This is somewhat greater than the amount of hominoid sequence data currently available, but not by even as much as an order of magnitude. The power of DNA hybridization methods is not comparable to that which would result from sequencing the entire single-copy fraction of the genome, because of the loss of resolution due to experimental measurement error. There is therefore the prospect that, if hybridization methods do not increase in power, sequence data could overtake hybridization data in the near future. However, the present calculation does not take into account either the degree of replication of hybridization values (which would favor hybridization data further) or the multispecies nature of the data set (which might perhaps favor sequence data, although this is not certain). It would be interesting to see a more detailed analysis allowing for both of these effects.

Acknowledgments. I am grateful to Alan Templeton for stimulating Sibley and Ahlquist to subject their data to a more complete statistical analysis. I thank Ruth Shaw and Elizabeth Thompson for helpful discussions concerning statistical methods. Ruth Shaw and William Hatheway are thanked for constructive comments on an earlier version of this paper. This work was supported by task agreement number DE-AT06-76EV71005 of contract number DE-AM06-76RL02225 between the US Department of Energy and the University of Washington.

References

Brown WM, Prager EM, Wang A, Wilson AC (1982) Mitochondrial DNA sequences of primates: tempo and mode of evolution. J Mol Evol 18:225–239

Dempster AP, Laird MN, Rubin DB (1977) Maximum likelihood from incomplete data via the EM algorithm. J Roy Statist Soc B 39:1–38

Farris JS (1981) Distance data in phylogenetic analysis. In: Funk VA, Brooks DR (eds) Advances in cladistics. Proceedings of the first meeting of the Willi Hennig Society. New York Botanical Garden, Bronx, pp 3–23

Farris JS (1985) Distance data revisited. Cladistics 1:67–85

Farris JS (1986) Distances and statistics. Cladistics 2:144–157

Felsenstein J (1983) Statistical inference of phylogenies. J Roy Statist Soc A 146:246–272

Felsenstein J (1984) Distance methods for inferring phylogenies: a justification. Evolution 38:16–24

Felsenstein J (1986) Distance methods: reply to Farris. Cladistics 2:130–143

Fitch WM, Margoliash E (1967) Construction of phylogenetic trees. Science 155:279–284

Jukes TH, Cantor CH (1969) Evolution of protein molecules. In: Munro HM (ed) Mammalian protein metabolism. Academic Press, New York, pp 21–123

Ruvolo M, Smith TF (1986) Phylogeny and DNA-DNA hybridization. Mol Biol Evol 3:285–289

Saitou N (1986) On the delta Q-test of Templeton. Mol Biol Evol 3:282–284

Sibley CG, Ahlquist JE (1984) The phylogeny of the hominoid primates, as indicated by DNA–DNA hybridization. J Mol Evol 20:2–15

Sibley CG, Ahlquist JE (1987) DNA hybridization evidence of hominoid phylogeny: results from an expanded data set. J Mol Evol 26:99–122

Templeton AR (1983) Phylogenetic inference from restriction endonuclease cleavage site maps with particular reference to the evolution of humans and the apes. Evolution 37:221–244

Templeton AR (1985) The phylogeny of the hominoid primates: a statistical analysis of the DNA–DNA hybridization data. Mol Biol Evol 2:420–433

Templeton AR (1986) Further comments on the statistical analysis of DNA–DNA hybridization data. Mol Biol Evol 3:290–295

Received January 22, 1987/Accepted April 6, 1987