Statistical Science
1987, Vol. 2, No. 2, 191–210

Statistical Analysis of Hominoid Molecular Evolution

Daniel Barry and J. A. Hartigan

Abstract. The core data of molecular biology consists of DNA sequences. We will show how DNA sequences may be used to infer the evolution of the primates, human, chimpanzee, ape, orangutan and gibbon. The underlying probability models are taken to be Markov processes on trees. Some dependencies along the sequence due to the genetic code are also considered.

Key words and phrases: Molecular evolution, Markov processes, DNA sequences, primate evolution.

1. INTRODUCTION

Since Darwin, man has been relegated from the angels to the apes, but exactly where in the apes? In 1960, the *Encyclopedia Britannica* places man outside the Pongidae, consisting of the gibbons, orangutans, chimpanzees and gorillas. In 1980, man is securely established in the group human, chimpanzee and gorilla, although there remains some controversy over his exact placement there. (See, however, Kluge (1983) who favors a human–orangutan grouping.)

The traditional data for inferring evolution come from morphology (form and structure) and paleontology (fossil remains). The new data, which are destined to resolve many long-standing problems in evolution and systematics, are based on the genetic macromolecules called deoxyribonucleic acid (DNA).

The DNA consists of a sequence of bases, either thymine (T), cytosine (C), adenine (A) or guanine (G), attached to a sugar-phosphate backbone and paired with a complementary sequence of the same bases attached to another such backbone. Thymine pairs with adenine and guanine with cytosine. Adenine and guanine are purines, composed of two carbon rings, and thymine and cytosine are pyrimidines, composed of one carbon ring. Human DNA is approximately 3×10^9 base pairs (bp) in length. The DNA generates ribonucleic acid (RNA) and proteins, which assist in the reproduction of the DNA. When an organism reproduces, it transmits DNA to its offspring, and this DNA determines the development of the new organism (see Alberts, Bray, Lewis, Raff, Roberts and Watson (1983) to discover the glorious details). *Scientific American* (October 1985) reviews recent advances.

In humans and apes, the DNA is divided between the nucleus and organelles outside the nucleus called mitochondria. The mitochondrial DNA is only 16,500 bp in length, and it is known to change somewhat more rapidly than nuclear DNA; it is therefore a good first candidate for comparative studies between close species. Each protein is a sequence of amino acids, of which there are 20; a *codon* of DNA is a sequence of three bases that translates to a particular amino acid. The code that translates DNA codons to amino acids in mitochondria is given in Figure 1. The nuclear code is slightly different. The four STOP codons specify the end of a segment of DNA that translates, via an RNA messenger strand, into protein. Note that the third position of a codon is degenerate, in the mitochondrial code, in that either purine or pyrimidine gives the same amino acid.

Until the late seventies, DNA sequences were mainly determined indirectly by sequencing the RNA or proteins that they generate. In 1977, two new techniques, named for the developers, Maxam-Gilbert and Sanger-Nicklen-Coulson, greatly speeded up the routine sequencing of DNA, and a flood of DNA sequences have since appeared (Table 1).

It is our intention to study a variety of molecular data to demonstrate a variety of statistical methods that might be used for inferring evolution. The particular evolutionary problem considered is the ancestry of the hominoids. The statistical literature is modest; the first probability models are due to Edwards and Cavalli-Sforza (1964); Neyman (1971) first presented models for nucleic acid sequences; and Felsenstein (1983) has an excellent review article on Markov process models.

Sections 2 through 4 review the evidence provided by chromosome comparisons, DNA hybridization, protein comparisons and cleavage maps. The remainder

Daniel Barry is Lecturer, Department of Statistics, University College, Cork, Ireland. J. A. Hartigan is Eugene Higgins Professor of Statistics, Department of Statistics, Yale University, Box 2179 Yale Station, New Haven, Connecticut 06520.

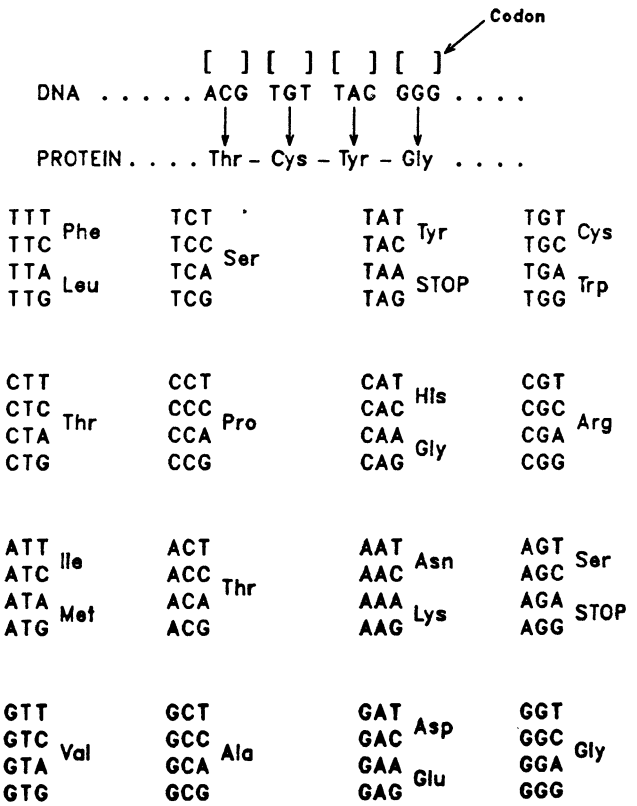

FIG. 1. *Mitochondrial genetic code. A triplet of nucleotides, or codon, translates to 1 of 20 amino acids according to the code. Note that the transitions in the third positions (changing the purine to a purine, or the pyrimidine to a pyrimidine) do not change the amino acid coded for.*

TABLE 1
DNA sequences in Gen bank (15 *Feb* 1984): 2,825,441 *bases at* 3,424 *loci*[a]

	Bases	Fraction of DNA
Human	279,837	0.00005
Mouse	257,207	0.00004
Rat	108,968	0.00002
Drosophila	95,825	
Sea urchin	38,682	
Yeast	105,594	
Escherichia coli	222,844	0.05
Phage lambda	49,789	1.00

[a] Typical fragment is 1 or 2 kb associated with a particular gene.

of the paper deals with techniques for analyzing DNA sequence data and applies these techniques to the hominoid DNA data. The data used is described in detail in Section 5. Graphical techniques for looking at single DNA sequences are developed in Section 6 along with Markov models to describe the dependence between sites indicated by the graphs. The concept of a silent site is described in Section 7 and the behavior of silent sites contrasted with that of non-silent or replacement sites. In Section 8 measures of distance between pairs of DNA sequences are described and a new measure proposed. Sections 9 through 11 include the specifications of a simple probability model for evolutionary change and the development of algorithms to fit this model to data using maximum likelihood. In Section 12 these algorithms are applied to the hominoid DNA data and we conclude that there is evidence (although not conclusive evidence) that human and chimpanzee branched most recently from the evolutionary tree of the hominoids.

2. LARGE SCALE MOLECULAR COMPARISONS

In this section, evidence based on viewing the DNA as a whole is considered. The DNA is divided into 23 pairs of chromosomes in humans, and 24 pairs in chimpanzees, gorillas and orangutans. Yunis and Prakash (1982) photographed G-banded chromosomes for humans, chimpanzees, gorillas and orangutans; G-banding is achieved by dyeing the chromosomes giving bands of different shades. Each human chromosome may be matched with a similar chromosome in each of the other species; all matched chromosomes are said to be homologous, descended from a common ancestral chromosome. One of the human chromosomes is homologous to two ape chromosomes. The main differences between homologous chromosomes made visible by G-banding are inversions, in which a segment of chromosome is reversed in direction. The chromosomes are about 10^8 bp in length, so differences between homologous segments less than 10^6 bp in length are not visible.

Yunis and Prakash judge that 13 of the human–chimpanzee chromosomes are identical, but only 9 of the human–gorilla and 8 of the human–orangutan are identical. This is evidence that human and chimpanzee branched most recently.

The most powerful global technique for comparing DNA is DNA hybridization, which has been used by Sibley and Ahlquist (1983) for an extensive revision of the systematics of birds. A distance between two pieces of DNA is measured by the propensity of single strands from each piece to reassociate as double strands:

1. Take a piece of double-stranded DNA, shear it by sonification into fragments about 500 bp in length, disassociate the fragments into single-stranded fragments by heating.

2. Cool and remove those fragments of the DNA that reassociate most quickly. These will be "repeated copy" sequences of DNA; perhaps half of the DNA is composed of sequences that repeat themselves, some sequences of length 10 repeated 10^7 times, some sequences of length 1000 repeated 10^4 times; these sequences are excluded from the comparison because they tend to reassociate with other copies in the same single strand.

3. A tracer piece is labeled radioactively, and the driver piece, in concentration 1000:1 to the tracer, is allowed to reassociate with it. The tracer piece is treated by steps 1 and 2, the driver by step 1. At this point almost all the tracer DNA will be associated with driver DNA; each fragment of tracer will be aligned, with some mismatches, with a fragment of driver. The mixture is heated until 50% of the tracer DNA is disassociated, and 50% is hybridized with the driver DNA—this temperature is called the $T_{50}H$, and the higher the $T_{50}H$ the more similar the two pieces of DNA, since dissimilar pieces will have many mismatches in the aligned fragments, and so the fragments will easily disassociate.

4. The $T_{50}H$ is determined first for tracer DNA and driver DNA from the same species, and then for tracer DNA from that species and driver DNA from another species. The difference between the two values of $T_{50}H$ is denoted $\Delta T_{50}H$ and measures distance between the tracer and driver species.

It is difficult to see, through the complex interaction of billions of fragments extensively treated, what kind of distance between molecules is measured by $\Delta T_{50}H$. From Sibley and Ahlquist (1983) $\Delta T_{50}H = 1°C$ corresponds to a 1% mismatch rate, based on experiments using synthetic polynucleotides of known composition. But there remains to be developed a model connecting the possible evolutionary changes—substitutions, insertions, deletions, inversions, to the kinetics of association of DNA fragments.

Average $\Delta T_{50}H$ distances between the hominoids are given in Table 2. Chimpanzee–human distances are smaller than the others, but the gorilla distances to human and chimpanzee are only slightly larger. Each of these averages is based on several separate ΔT_{50} computations, so that it is possible to use standard analysis of variance (one-way classification) techniques to evaluate various hypotheses about the underlying distances. It might be expected that the errors are somewhat larger in estimating distances between the more distant species, but they did not seem to be large enough to affect the analysis. It might also be argued that the analysis should be directed at the original T_{50} values rather than the ΔT_{50} values; a more sophisticated analysis should allow for different errors associated with each driver and tracer preparation. We are first interested in detecting asymmetry in the matrix; we hope to find that the tracer–driver distances are not significantly different from the corresponding driver–tracer distances. There are two additive models that may be associated with a given evolutionary tree.

1. *Asynchronous models*, in which the distance between two species is the sum of the "amounts of evolutionary change" in each of the links on the path connecting the two species. See, for example, Cavalli-Sforza and Edwards (1967).

2. *Synchronous models*, in which the distances satisfy (1), but also the total evolutionary change from the root of the tree to all present day species is the same; the term synchronous is used because the models follow from the assumption that the amount of change in a given time period is the same across all lineages, although not necessarily the same in different time periods of the same length. In this case the distances form an ultrametric (Hartigan, 1967).

Denote by HCG the models in which chimpanzee, human and gorilla have the same branch point, by HC the models in which human and chimpanzee have the most recent branch point and by CG the models in which gorilla and chimpanzee have the most recent branch point. The letters A and S are used to distinguish asynchronous from synchronous models. Each model represents the $\Delta T_{50}H$ distances as a sum of parameters (see Figure 2). The parameters are estimated by least squares, and the models evaluated by the residual sum of squares computed from the original 147 distances of Sibley and Ahlquist (1983). The two bifurcation models are compared to the trifurcation model, and the synchronous models are compared to the corresponding asynchronous models in Table 3.

If we could be sure of synchronous evolution, the evidence for HC would be compelling (the F value is 27 for HC against trifurcation). Sibley and Ahlquist

TABLE 2
Average ΔT_{50} for hominoids

Tracer	Driver							
	Pan	Pt	Hs	Gg	Pp	Hl	Hsyn	
Pan paniscus		.78	1.77	2.38	3.87	5.58	[a]	Chimpanzees
Pan troglodytes	.50		1.72	2.12	3.60	4.98	[a]	
Homo sapiens	1.77	1.83		2.54	3.67	5.10	4.80	Human
Gorilla gorilla	2.00	2.14	2.24		3.90	5.43	[a]	Gorilla
Pongo pigmaeus	3.55	3.72	3.60	3.70		5.33	5.09	Orangutan
Hylobates lar	5.90	5.16	5.28	5.28	4.98		2.15	Gibbons
Hylobates syndactulus	[a]	[a]	[a]	[a]	[a]	[a]		

[a] No data available.

(1984) base their analysis on synchronous evolution, which they call the uniform average rate hypothesis. The data suggest some asynchronous behavior (the F for AHC against SHC is 2.98), and this reduces the strength of the evidence for HC, but HC remains clearly significant against HCG and clearly preferred to CG.

Finally, we may compare the asynchronous human–chimpanzee model with the network model (NET) in which distances are unconstrained, except that they be symmetrical. Then $F_{7,129} = 2.03$ has a tail probability of 0.05. Against the synchronous model, $F_{12,129} = 2.50$ has tail probability 0.001. Thus, we can barely be satisfied with fitting an asynchronous CH model, but have too much error for the synchronous CH model. The network model would be superior to the tree models if convergent evolution occurs: different lineages become more similar after first diverging.

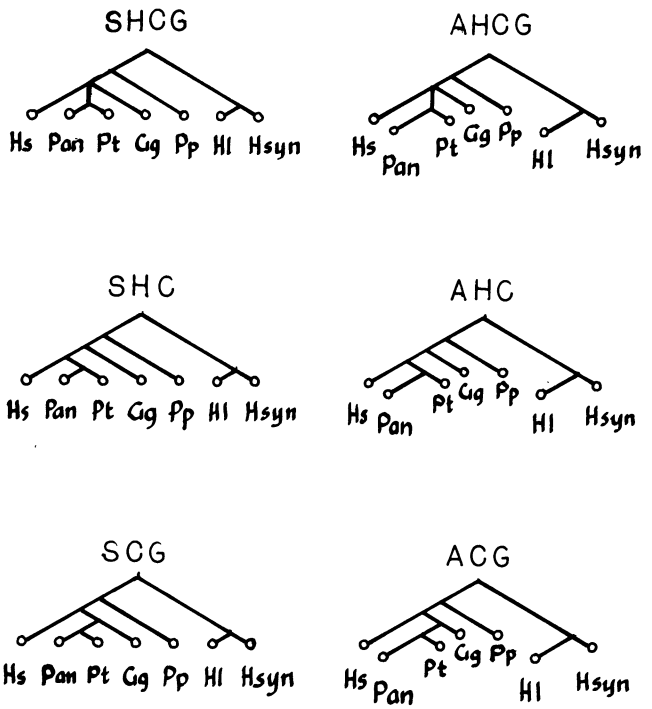

FIG. 2. *Synchronous and asynchronous models for hominoid evolution.* HCG *denotes human, chimpanzee and gorilla branching at the same time;* HC *denotes human and chimpanzee branching more recently than gorilla; and* CG *denotes chimpanzee and gorilla branching more recently. Evolution may be synchronous* (S), *where the rate of change is the same in all lines, or asynchronous* (A) *where the rate of change varies in different lines.*

3. PROTEIN EVIDENCE

A typical protein is 100 amino acids long, generated by 300 bp of DNA. Since proteins have functions of varying importance to the survival of the DNA, there are widely varying rates of change for different proteins. Until the late 70s data of this type were the main source of information about DNA (see Dayhoff, 1978). For studying closely related species such as the hominoids, it is necessary to look only at fast changing proteins.

In Table 4 are given the hominoid sequences for fibrinopeptides A and B, proteins that are linked to form fibrinogen. These are among the fastest evolving proteins; fibrinopeptide A changes only in the third position, suggesting the grouping (human, chimpanzee, gorilla, orangutan); the ancestral value threonine changed after the branch point between *Hylobates* and pongids to serine—such an event, where two species have in common a character differing from an ancestral character, is called a *synapomorphy.* Two synapomorphies appear in fibrinopeptide B, and in addition there is a deletion in the gibbon sequence. (In long protein sequences, the possibility of many insertions and deletions creates a different realignment problem for which a number of algorithms have been developed—these are variations of the Erdös and Szekeres (1935) method for finding the longest increasing subsequence in a sequence of integers; see Sankoff and Kruskal, 1983.)

TABLE 3
Analysis of variance of Sibley-Ahlquist data[a]

Model	Residual sum of squares	d.f.	Including model	F value
No constraints	8.452	114		
Symmetry	9.832	129	No constraints	1.24
AHC	10.917	136	Symmetry	2.03
ACG	11.577	136	Symmetry	3.27
AHCG	11.580	137	AHC or ACG	8.26 or .03
SHC	12.115	141	AHC	2.98
SCG	14.098	141	ACG	5.91
SHCG	14.464	140	SHC or SCG	27.34 or 3.66

[a] No constraints, distances may be asymmetrical; symmetry, distances symmetrical; AHCG, asynchronous, chimpanzee–human–gorilla branch simultaneously; AHC, asynchronous, chimpanzee–human split latest; ACG, asynchronous, chimpanzee–gorilla split latest; SHCG, synchronous, chimpanzee–human–gorilla branch simultaneously; SHC, synchronous, chimpanzee–human split latest; SCG, synchronous, chimpanzee–gorilla split latest.

The synapomorphies in fibrinopeptide B suggest the branch human–chimpanzee–gorilla. Note that the ambiguous subsequence BBBZZ is probably NDNEE. Goodman et al. (1982) find seven positions in β-hemoglobin, fibrinopeptides and myoglobin that separate human–chimpanzee–gorilla from gibbon–orangutan; they find one position in α-hemoglobin and four positions in carbonic anhydrase I separating human–chimpanzee from gorilla–orangutan. Thus, there are five synapomorphies in favor of human–chimpanzee.

4. CLEAVAGE MAPS

Ferris, Wilson and Brown (1981) studied mitochondrial DNA in hominoids using cleavage maps. Mitochondria are organelles in eucaryotic cells that carry their own DNA, and that reproduce by mitosis, inheritance being from the mother. Thus, mitochondrial evolution is more similar to bacterial evolution than to evolution of nuclear DNA. Mitochondrial DNA is a circle of about 16,500 bp in mammals; it evolves rapidly and so is especially suited to studying hominoid evolution.

Restriction enzymes cleave the DNA at specific sequences, usually four or six in length, and usually *palindromic*—for example the enzyme HindIII cleaves sequences AAGCTT; the complementary sequence is TTCGAA, which is the reverse of the original sequence. Restriction enzymes were invented by bacteria as weapons to hurl at invading viruses, cutting them into harmless little pieces. (Of course the bacteria have to arm their vulnerable segments, and the viruses try to acquire the same armor.)

Ferris, Wilson and Brown studied the sites in hominoid mitochondrial DNA where 19 restriction enzymes cleave, about 50 sites in all; we see the behavior of the DNA at $300 = 6 \times 50$ bp. Synapomorphies are sites where two sets of two species have different restriction behavior; Table 5 favors gorilla–chimp (seven synapomorphies) over chimp–human (three synapomorphies).

TABLE 4
Hominoid fibrinopeptides[a]

	Fibrinopeptide A	Fibrinopeptide B
Human	ADSGEGDFLAEGGGVR	ZGVNDNEEGFFSAR
Chimpanzee	ADSGEGDFLAEGGGVR	ZGVNDNEEGFFSAR
Gorilla	ADSGEGDFLAEGGGVR	ZGVNDNEEGFFSAR
Orangutan	ADSGEGDFLAEGGGVR	ZGVBBBZZGLFGAR
Siamang	ADTGEGDFLAEGGGVR	ZGVBBBZZGLFGAR
Gibbon	ADTGEGEFLAEGGGVR	ZGVBBBZGLFGAR
	1	2 3 4

[a] 1, 3, 4: A synapomorphy, a set of species sharing a character differing from the ancestral character; 2: a deletion in gibbon, requiring realignment. A, alanine; B, aspartic acid or asparagine; D, aspartic acid; E, glutamic acid; F, phenylalanine; G, glycine; L, leucine; N, asparagine; R, arginine; S, serine; T, threonine; V, valine; Z, glutamic acid or glutamine.

5. MITOCHONDRIAL DNA

Brown, Prager, Wang and Wilson (1982) looked at complete mitochondrial DNA sequences for five hominoids, at 896 sites between two HindIII cleavage sites. Chimpanzee had a HindIII cleavage site at position 263, so it was necessary to combine fragments of length 263 and 633 to obtain the chimpanzee sequence. The orangutan had a deletion at site 562.

Brown, Prager, Wang and Wilson (1982) compute for each possible evolutionary tree the minimum number of substitutions required to reproduce the observed sequences for the five hominoids. Last split gorilla–chimpanzee requires 145 substitutions, human–chimpanzee 147 substitutions and human–gorilla 148 substitutions. They conclude that the data favors gorilla–chimpanzee but does not rule out the other splits. Our analyses to follow show that the data favor, slightly, a human–chimpanzee branch.

6. MARKOV MODELS ALONG HUMAN DNA SEQUENCES

Many of the probability models used in constructing evolutionary trees assume that the sites on the DNA molecule are independent and identically distributed (iid) over the set of bases {A, C, G, T}. However, some patterns are observable in DNA sequences. For example, purines tend to follow purines and pyrimidines tend to follow pyrimidines. Certain subsequences tend to occur more frequently than others

TABLE 5
Synapomorphies at cleavage sites in mitochondrial DNA[a]

	Human	Chimpanzee	Gorilla	Orangutan	Gibbon
Human		3	1	1	2
Chimpanzee	0		6	1	2
Gorilla	0	1		1	1
Orangutan	2	1	0		4
Gibbon	0	1	2	3	

[a] Above diagonal, sites cleaved only at the given two species; below diagonal, sites not cleaved only at the given two species.

and, as will be seen in Section 7, some sites are more prone to change than others.

It is possible to represent a DNA sequence graphically using the following two plots:

(1) Let $X_i = +1$ if position i is C or T,
-1 if position i is A or G.
(2) Let $Y_i = +1$ if position i is C or G,
-1 if position i is A or T.
Let $SX_i = \sum_{j=1}^{i} X_j$. Plot SX_i versus i.
Let $SY_i = \sum_{j=1}^{i} Y_j$. Plot SY_i versus i.

The first plot looks at the preponderance of pyrimidines over purines and the second looks at the preponderance of one pair of complementary bases over the other. Figure 3 shows these plots for the human mitochondrial DNA sequences. Long series of pyrimidines occur starting approximately at positions 350 and 650. One would not expect to see such phenomena in an iid sequence. The complementary base plot reveals less of a departure from the iid model.

The adequacy of the iid model for representing long range patterns in a DNA sequence under study can be checked by comparing the above graphs calculated using the real sequence with graphs calculated using sequences simulated from the iid model.

We considered a generalization of the iid model to a Markov model. The probability of the whole sequence is determined by the conditional distribution of the next base given the preceding sequence of bases. It is assumed that this conditional distribution is determined by just a few different subsequences which we call the effectives. If there are no effectives then the DNA is just a sequence of independent and identically distributed assignments of bases from the set {A, C, G, T}. We shall describe this case by saying that there is one effective which we shall denote by ·. The long nonrandom patterns observable in some parts of the DNA may make this an inappropriate probability model. For an actual sequence it is necessary to build a set of effectives, ones with different conditional distributions for the next base, to maximize the probability of the observed sequence.

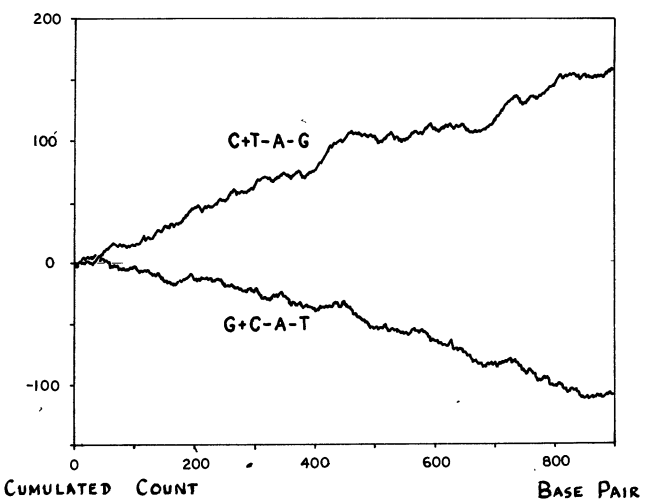

FIG. 3. GATC-TAGC *plots of 896 bp of human mitochondrial DNA. The top plot is the cumulative count of the number of* Cs *and* Ts *less the number of* Gs *and* As, *that is the number of pyrmidines less the number of purines. The sequence is pyrimidine-rich, especially in sequences near 400 and 700 bp. The lower plot is the cumulative number of* Gs *and* Cs *less the number of* As *and* Ts, *the difference between the two complementary pairs. There are more* As *and* Ts, *but the distribution over the sequence is apparently random. A third plot* G + T − C − A *would complement these and make possible complete recovery of the sequence.*

Given a set of effectives and their associated conditional distributions, it is possible to generate random sequences having the same distribution. These may be examined to see how well they reflect the large scale behavior of observed DNA strands.

Suppose we are given a set of effectives E. Each site on the DNA is assigned to the longest effective immediately preceding it. The conditional probabilities for each effective can be calculated from the frequencies at sites assigned to it. The log likelihood of the observed sequence $\mathbf{x} = (x_1 x_2 \cdots x_n)$ is then $\sum \log p(x_i)$, where $p(x_i)$ is the probability of the ith element calculated using the conditional distribution of the effective to which it is assigned.

We describe an algorithm which may be used to find all effectives up to a maximum length, MAX.

1. We begin with only · as an effective and calculate the log likelihood as above.
2. $L = 1$.
3. Find the increase in log likelihood obtained by adding each subsequence of length L to the set of effectives, E.
4. Add to E any subsequence for which the increase is bigger than some prechosen limit ADD + L.
5. Find the drop in log likelihood obtained by dropping each effective from E.
6. If the smallest such drop is smaller than a prechosen limit DROP + L, drop that effective from E and go to 5. (We add L to ADD and DROP to allow for the increase in number of available effectives as L increases.) If the smallest drop is larger than DROP + L, $L = L + 1$.
7. If $L <$ MAX go to 3; otherwise STOP.

When this algorithm (with ADD = DROP = 2, MAX $L = 6$) was applied to the human mitochondrial DNA sequences only two effectives were found — · and the effective C of length 1. The increase in log likelihood for this model as compared to the independence model was $-1172.8 - (-1179.6) = 6.8$. This has significance level less than 0.005 when compared with a $\frac{1}{2}\chi_3^2$ variable; some adjustment must be made to allow for selecting the best of four effectives of length 1. The conditional probabilities are as in

TABLE 6
Markov models along DNA

Effective	A	C	G	T	No. of occurrences
(i) iid model					
.	·30	·33	·11	·26	896
(ii) Markov Model I					
.	·31	·33	·13	·23	599
C	·31	·33	·06	·30	297
(iii) Markov Model II					
.	·30	·34	·13	·23	580
C	·34	·26	·07	·33	187
CC	·19	·39	·00	·42	54
ACC	·29	·46	·11	·14	28
CCG	·50	·00	·50	·00	4
GTA	1·00	·00	·00	·00	4
CCCC	·44	·44	·11	·11	16
TAAC	·25	·67	·00	·08	12
TCAT	·36	·00	·28	·36	11

Table 6. Clearly more T's and fewer G's tend to follow C's in the sequence. This is an example of pyrimidines following each other.

When the algorithm was applied with a less conservative ADD = DROP = 0.5, the set of effectives found is listed in Table 6.

The increase in log likelihood for this model as compared to the independence model was $-1141.9 - (-1179.6) = 37.7$, which has a nominal significance level less than 0.005 when compared with a $\frac{1}{2}\chi^2_{24}$ variable. The significance level is untrustworthy since we searched over a huge class of possible effectives to reach the final model. The first Markov model may be more appropriate in this case.

7. REPLACEMENT SITES AND SILENT SITES

In Figure 4, the pattern of substitutions between human and gibbon is given in six strips. Each strip is composed of *blanks or circles* (representing substitutions) in columns of three; the first column corresponds to sites 1, 2, 3, the second to 4, 5, 6 and so on. It is evident that substitutions occur much more frequently in the top row of the strips in the first part of the data and in the bottom row in the later part of the data.

The mitochondrial segment is composed of part of an unidentified reading frame, URF4, coding for some unidentified protein; three short segments of about 80 bp coding for transfer RNA for histidine, serine and leucine; and part of an unidentified reading frame, URF5. From the genetic code for mitochondrial DNA, the value of the site in the third position of the codon may be either purine or pyrimidine without affecting the amino acid coded for. A change from one purine to the other or from one pyrimidine to the other is said to be a *transition*; other changes are said to be

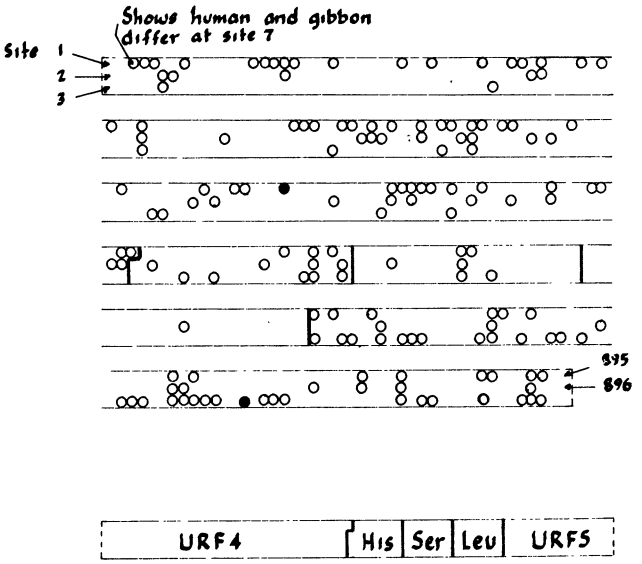

FIG. 4. *Sites where humans and gibbons differ. The 896 sites are arranged in triplets. Circles mark the sites where humans and gibbons differ. The three interior segments code for the transfer RNAs for histidine, serine and leucine; note that the leucine transfer RNA has only one site of change. The outer segments code for unidentified proteins; the number of changes in the top row of the first segment and the bottom row of the last is greater than in the other two rows combined; these are the third codon changes. The only third codon changes that cause a change in the protein coded for are indicated by two filled circles; thus, third codon changes are nearly all silent. Conversely almost all first and second codon changes affect the protein.*

transversions. Thus, a transition in the third position has no effect on the protein: the third position is said to be silent. (Even transversions have no effect in the third position of 6 of the 20 codons.) Changes in the other sites cause replacement of an amino acid in the protein—these are said to be replacement sites. (The code for nuclear DNA is slightly different and some transitions in third position do effect the protein translated to.)

Here the top row of the strip at the beginning of the segment and the bottom row at the bottom are silent; the transfer RNAs appear in the middle of the segment; the last transfer RNA is significantly well-conserved. Empirical evidence suggests treating silent sites differently from the others—indeed, since they do not affect the protein, they offer a selection-neutral history of change.

8. MEASURES OF DISTANCE BETWEEN SPECIES

Measures of distance between species are an important first step in constructing evolutionary trees. Numerous measures are proposed and we will use the mitochondrial DNA data to examine the usefulness of each.

A simple measure of distance is the number of sites at which the two species differ. These distances are shown in Table 7 for the five hominoid species. The chimpanzee and human seem closest, whereas the gibbon appears quite different from each of the other four species.

A synapomorphy exists when a pair of species has the same value at a site differing from the ancestral value at that site. It is usually necessary to infer the ancestral value. A synapomorphy for two or more species is taken as evidence that they have a common ancestor differing from the ancestor for all species. Consider two examples:

	C	Gi	Go	O	H
1.	C	C	T	T	T
2.	T	G	A	A	T

Example 1 shows a synapomorphy between C and Gi, T being taken as the ancestral value. Example 2 is more complicated. We scored this as a synapomorphy between C and H arguing that the distant Gi changed from a G to an A for the group {C, Go, O, H} and later C and H both substituted a T for A. In general in dealing with 2 − 2 − 1 splits we assumed that the gibbon is most distant followed by the orangutan and assigned synapomorphies accordingly. These assumptions do not bias the comparison between C, H and G. Such splits only occurred eight times. Table 8 shows the number of synapomorphies between each pair of species. Large values indicate closeness. Here orangutan and gibbon seem close, but the chimpanzee seems about equidistant from human and gorilla. This picture is quite different from that obtained using raw differences.

Table 9 shows separate observed substitution rates for silent, replacement and transfer sites. Clearly the substitution rates for silent sites are much higher. However, the rates are not proportional to the replacement rates; silent sites show much smaller relative distances between species. The distances calculated using replacement and transfer sites are similar to one another and lead to conclusions similar to those arrived at by consideration of raw differences between complete sequences.

TABLE 7
Number of differences between species

	H	C	Go	O	Gi
H		79	92	144	162
C			95	154	169
Go				150	169
O					169
Gi					

TABLE 8
Number of synapomorphies between species

	H	C	Go	O	Gi
H		14	8	2	3
C			11	3	6
Go				6	9
O					28
Gi					

TABLE 9
Observed substitutions

	C	Go	O	Gi
All ($n = 896$)				
H	.09	.10	.16	.18
C		.11	.17	.19
Go			.17	.19
O				.19
Silent ($n = 232$)				
H	.23	.24	.32	.36
C		.27	.34	.39
Go			.30	.37
O				.40
Replacement ($n = 465$)				
H	.04	.05	.11	.13
C		.05	.11	.12
Go			.13	.13
O				.12
Transfer ($n = 198$)				
H	.05	.06	.09	.11
C		.07	.11	.12
Go			.09	.11
O				.10

The observed number of substitutions underestimates the total number occurring between two species since a series of substitutions at the same site produces the same final effect as just one substitution. It is therefore necessary to adjust the observed substitution rates to allow for this possibility.

Let P be the matrix of transition probabilities between two species (P is a 4×4 matrix where, for example, P_{AC} is the probability of a C in the second species given an A in the first). It is proposed to estimate the substitution rate by

$$r = -\tfrac{1}{4} \log\{\det(P)\}.$$

This choice can be justified as follows. Suppose

$$P = \prod_{\alpha} P_\alpha$$

where $\{P_\alpha\}$ are infinitesimal transition matrices. Then

$$\log\{\det(P)\} = \sum_{\alpha} \log\{\det(P_\alpha)\}.$$

Let $P_\alpha = (P_{ij})$. Then

$$P_{ij} = \begin{cases} 1 + a_{ij}, & i \neq j, \\ a_{ij}, & i \neq j, \end{cases}$$

where the (a_{ij}) are small. Hence,

$$\det(P_\alpha) \approx 1 + a_{11} + a_{22} + a_{33} + a_{44} \Rightarrow -\log\{\det(P_\alpha)\}$$
$$\approx -(a_{11} + a_{22} + a_{33} + a_{44})$$
$$= (1 - P_{11}) + (1 - P_{22}) + (1 - P_{33}) + (1 - P_{44})$$
$$= \sum_{i=1}^{4} P[\text{change} \mid i]$$

(where i is the value at start of increment α)

$$= \sum_{i=1}^{4} P[\text{change from } i]/P(i).$$

So changes from i are weighted by $1/P(i)$. The average value of $1/P(i)$ is 4 and so $-\frac{1}{4}\log\{\det(P_\alpha)\}$ estimates the number of changes in increment α, where a change from i is weighted by $1/P(i)$. There is no assumption that the infinitesimal transition matrices are the same or that the process is stationary—in general $-\frac{1}{4}\log\det P$ estimates the total number of changes between two species, with each change weighted inversely by the probability of the base changed from.

Table 10 shows the distances calculated using this measure. The distance from species A to B is not necessarily equal to the distance from B to A. Here they are very close and have been averaged. Again silent sites show much higher substitution rates than replacement or transfer sites although again relative distances are smaller for silent sites. In some cases, the determinant was negative. This may be interpreted to mean that so much change has occurred that it is not possible to estimate distance. The estimated number of substitutions is about three times the observed number for the silent sites, and about the same for the replacement sites. In all cases human–chimpanzee are the closest pair but the differences are not statistically significant.

TABLE 10
Estimated substitutions (log det)

	C	Go	O	Gi
All				
H	.10	.12	.20	.23
C		.13	.21	.24
Go			.21	.23
O				.24
Silent				
H	.66	.77	.91	1.26
C		.87	1.47	[a]
Go			1.20	1.58
O				[a]
Replacement				
H	.04	.06	.14	.15
C		.05	.14	.14
Go			.17	.16
O				.14
Transfer				
H	.05	.06	.11	.11
C		.07	.14	.14
Go			.11	.13
O				.11

[a] Negative determinant.

Felsenstein (1983) considers constant rate models for the transition probabilities. Let $Q = (q_{ij})$ be a 4×4 matrix where $q_{ij}\delta t$, $i \neq j$, is the probability of state j at time $t + \delta t$ given state i at time t; $q_{ii} = -\sum_{j \neq i} q_{ij}$. Then the transition matrix over a time period t, P_t, is given by

$$P_t = e^{Qt} = I + Qt + \tfrac{1}{2}Q^2t^2 + \cdots.$$

In order to determine the expected number of substitutions, it is necessary to assume that the process is *reversible*, that is the transition matrix in going from 0 to t equals the transition matrix in going from t to 0. Let m_i be the stationary probabilities; then $\sum m_i q_{ij} = 0$ and the process is reversible if and only if $m_i q_{ij} = m_j q_{ji}$.

Assume without loss of generality that the total time elapsed between two species is $t = 1$, so that $P = e^Q$. In this model we can estimate the substitution rate as follows:

expected no. of substitutions in time δt

$$= \sum_{i=1}^{4} m_i \sum_{j \neq i} q_{ij}\delta t$$
$$= -\sum_{i=1}^{4} m_i q_{ii}\delta t.$$

Hence,

expected no. of substitutions between species

$$= -\sum_{i=1}^{4} \sum_{\delta t} m_i q_{ii}\delta t$$
$$= -\sum_{i=1}^{4} m_i q_{ii}.$$

Let $H = (h_{ij})$ be the observed joint probability matrix between two species. To calculate the above quantity we first symmetrize the data by creating a symmetric joint probability matrix H^*,

$$h^*_{ij} = \tfrac{1}{2}(h_{ij} + h_{ji}).$$

From H^* we calculate m_i and a transition matrix P. Then

$$Q = \log(P)$$
$$= \log(I - (I - P))$$
$$= -(I - P) - \tfrac{1}{2}(I - P)^2 - \tfrac{1}{3}(I - P)^3 - \cdots.$$

This series converges to Q if the eigenvalues of $I - P$ are less than one in absolute value.

TABLE 11
Estimated substitutions at constant rate

	C	Go	O	Gi
All				
H	.10	.11	.19	.22
C		.12	.21	.23
Go			.20	.23
O				.23
Silent				
H	.38	.45	.67	.78
C		.53	.80	[a]
Go			.84	1.11
O				[a]
Replacement				
H	.04	.05	.12	.14
C		.05	.13	.13
Go			.15	.15
O				.14
Transfer				
H	.05	.06	.10	.11
C		.07	.13	.14
Go			.10	.13
O				.11

[a] No Q exists.

TABLE 12
Human–gibbon joint distribution

Human	Gibbon			
	A	G	T	C
A	.269	.023	.003	.009
G	.019	.085	.001	.002
T	.008	.001	.203	.045
C	.020	.006	.042	.264

Table 11 shows these distances calculated for the hominoid data. Here [a] means that the series defining Q did not converge. The results are similar to those obtained using $-\frac{1}{4}\log\{\det(P)\}$ as a measure of distance; however, the silent sites are estimated to be 50% further apart using $-\frac{1}{4}\log\det P$.

Table 12 shows the human–gibbon joint distributions. Clearly the symmetrization process has only a small effect in this case.

9. CONSTRUCTING EVOLUTIONARY TREES

Consider the construction of an evolutionary tree based on data from five species. A general tree structure appropriate to this case is shown in Figure 5.

The tree consists of eight nodes labeled 0 through 7. Nodes 0, 4, 5, 6 and 7 correspond to the five species making up the available data. No direct data is available for nodes 1, 2 and 3.

There are N sites, and it is supposed that a sequence of bases occurs for each node over the N sites. Each base is one of four possible nucleotides: adenine, guanine, cytosine, thymine.

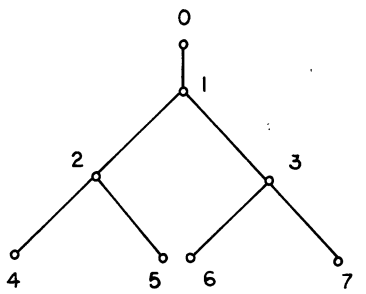

FIG. 5. *Evolutionary tree for five species. The observed species are at nodes* 0, 4, 5, 6, 7. *No data is available at internal nodes* 1, 2, 3.

We make the following assumptions:

1. At each site, the eight bases at the eight nodes are distributed identically as, and independently of, sets of eight bases at the other sites.
2. At each site, the bases are distributed as a Markov process: given the base at an internal node, the three sets of bases that remain connected when the internal node is removed are distributed independently of each other. For example, given the base at node 2, the bases at {0, 1, 3, 6, 7}, {4} and {5} are independent.

This model is more general than the one discussed in Felsenstein (1983); there each link in the tree has transition matrix $P_j = e^{Qv_j}$, where v_j is a parameter indicating length of link; here, P_j is arbitrary.

For each species we have a sequence of length N made up of the letters A, G, C and T. The five species can be allocated to the end nodes (i.e., nodes 0, 4, 5, 6 and 7) in 15 different ways. We proceed by calculating the likelihood for each of the resulting trees.

Write $\mathbf{X}_k = \{x_{ki}: i = 1, 2, \cdots, N\}$ for the sequence at node k. Then the likelihood of the tree may be written as

$$L = \prod_{i=1}^{N} P_0(x_{0i})P_1(x_{0i}, x_{1i}) \times P_2(x_{1i}, x_{2i})P_3(x_{1i}, x_{3i})P_4(x_{2i}, x_{4i}) \times P_5(x_{2i}, x_{5i})P_6(x_{3i}, x_{6i})P_7(x_{3i}, x_{7i}),$$

where $P_j(x_1, x_2)$ is the probability that a site is x_2 at node j given that it is x_1 at the beginning of the link leading into node j. P_0 is the marginal probability at node 0. (Node 0 is chosen as the root node, but any node could be so chosen.)

The unknowns in the above expression for the likelihood are $P_0, P_1, P_2, \cdots, P_7, \mathbf{X}_1, \mathbf{X}_2, \mathbf{X}_3$. In handling these unknowns two different approaches were considered.

1. *Most parsimonious likelihood.* All the unknowns, including the unknown values at internal nodes, are estimated by maximum likelihood and the likelihood of the tree calculated using these estimated values. We call this technique most parsimonious because the values of internal nodes are usually assigned to agree

as much as possible with neighboring nodes. It is therefore similar to the maximum parsimony fitting techniques in which values are assigned to internal nodes to minimize change between neighboring nodes (Felsenstein, 1983). This technique is easier to apply than the maximum average technique, but based on Felsenstein (1983) we expect the estimates of the transition matrices to be inconsistent as the number of sites approaches infinity.

2. *Maximum average likelihood.* In this approach we sum over the possible values of x_{1i}, x_{2i}, x_{3i} before taking products to get

$$L_1 = \prod_{i=1}^{N} \sum_{x_{i1}} \sum_{x_{i2}} \sum_{x_{i3}} P_0(x_{0i})P_1(x_{0i}, x_{1i})P_2(x_{1i}, x_{2i})P_3(x_{1i}, x_{3i}) \times P_4(x_{2i}, x_{4i})P_5(x_{2i}, x_{5i})P_6(x_{3i}, x_{6i})P_7(x_{3i}, x_{7i}).$$

In this expression only P_0, P_1, P_2, $\cdots$, P_7 are left as unknowns. These may be estimated by the values which maximize L_1 and the likelihood of the tree is taken to be the value of L_1 obtained using these estimates.

It is desirable to associate a distance measuring evolutionary change with each link in the tree.

It is clear how this should be done in the most parsimonious likelihood case. We can write

$$L = L_0L_1L_2 \cdots L_7$$

where

$$L_0 = \prod_{i=1}^{N} P_0(x_{0i}),$$

$$L_1 = \prod_{i=1}^{N} P_1(x_{0i}, x_{1i}),$$

$L_2, L_3, \cdots, L_7$ similarly.

We associate the distance $-\log L_j$ with the link leading into node j. These distances become bigger as the conditional probabilities become more spread out indicating more change along that link.

In the average likelihood case, the measure of distance used is $-\frac{1}{4}\log[\det(P_j)]$ where $\det(P_j)$ is the determinant of the transition probability matrix along the link leading into node j.

The proposed probability model is not identifiable. Consider the tree shown in Figure 5. Here nodes 0, 4, 5, 6 and 7 correspond to observed sequences. We can relabel the bases A, C, G, T at position 1 and positions 2, 3 and change the link transition matrices accordingly without changing the final likelihood. So, many sets of transition matrices lead to the same likelihood. If we insist that no change has higher conditional probability than any particular change, the relabelling of internal nodes will be prevented.

Neyman (1971) and Felsenstein (1981) have proposed parametric forms for the transition probabilities $P_1, P_2, \cdots, P_7$. Felsenstein assumes that in a small interval of time of length dt there is a probability udt that the current base at the site is replaced. If a base is replaced its replacement is A, C, G or T with probabilities π_1, π_2, π_3 or π_4. Let $P_{ij}(t)$ be the probability that a site which is initially in state i will be in state j after t units of time have elapsed. The above assumptions lead to

$$P_{ij}(t) = \begin{cases} e^{-ut} + (1 - e^{-ut})\pi_j, & i = j, \\ (1 - e^{-ut})\pi_j, & i \neq j. \end{cases}$$

In our notation

$$P_j(x_1, x_2) = \begin{cases} e^{-kj} + (1 - e^{-kj})P_0(x_2), & \text{if } x_1 = x_2, \\ (1 - e^{-kj})P_0(x_2), & \text{if } x_1 \neq x_2. \end{cases}$$

P_0 is assumed known and Felsenstein estimates $k_1, k_2, \cdots, k_7$ by maximum likelihood. Neyman's model is essentially the same except that he suggests $P_0(x_1) = \frac{1}{4}$ for all x_1.

Our model imposes no structure on the transition probabilities—at a cost of 12 parameters for each link instead of Felsenstein's one.

10. THE ALGORITHMS

1. *Most parsimonious likelihood case (MPL).*

(i) Initialize $\mathbf{X}_1$ by setting

$$x_{1i} = \text{mode}\{x_{0i}, x_{4i}, x_{5i}, x_{6i}, x_{7i}\}, \quad i = 1, 2, \cdots, N.$$

(ii) Initialize $\mathbf{X}_2$ and $\mathbf{X}_3$ by setting

$$x_{2i} = \text{mode}\{x_{1i}, x_{4i}, x_{5i}\},$$

$$x_{3i} = \text{mode}\{x_{1i}, x_{6i}, x_{7i}\}, \quad i = 1, 2, \cdots, N.$$

(iii) Estimate $P_j(x_1, x_2)$ by the proportion of times a transition from x_1 to x_2 occurs along the link leading into node j. Estimate $P_0(x_1)$ by the proportion of times x_1 occurs in $\mathbf{X}_0$. Calculate the likelihood L.

(iv) Update $\mathbf{X}_2$, $\mathbf{X}_3$ and $\mathbf{X}_1$ by choosing

x_{2i} to maximize $P_2(x_{1i}, x_{2i})P_4(x_{2i}, x_{4i})P_5(x_{2i}, x_{5i})$,

x_{3i} to maximize $P_3(x_{1i}, x_{3i})P_6(x_{3i}, x_{6i})P_7(x_{3i}, x_{7i})$,

x_{1i} to maximize $P_1(x_{0i}, x_{1i})P_2(x_{1i}, x_{2i})P_3(x_{1i}, x_{3i})$.

(v) Estimate the probabilities as in (iii). Calculate the likelihood.

(vi) If the increase in likelihood is large enough, repeat (iv) and (v). Otherwise stop.

2. *Maximum average likelihood (MAL).* The expression defining L_1 is the probability of the observed data. Write $Q_1(x_0, x_1)$ for the joint probability along the link leading into node 1. Similarly define $Q_2, Q_3, \cdots, Q_7$. Then we can write

$$L_1 = \prod_{i=1}^{N} \sum_{x_1} Q_1(x_{0i}, x_1)P(R_i \mid x_1)$$

where

$$P(R_i \mid x_1) = P(x_{4i}, x_{5i}, x_{6i}, x_{7i} \mid x_1)$$
$$= \sum_{x_2} \sum_{x_3} P_2(x_1, x_2)(P_4(x_2, x_{4i})P_5(x_2, x_{5i})$$
$$\times P_3(x_1, x_3)P_6(x_3, x_{6i})P_7(x_3, x_{7i}).$$

Since $\sum_{x_0} \sum_{x_1} Q_1(x_0, x_1) = 1$, we choose $Q_1(x_0, x_1)$ to maximize

$$\log L_1 + \lambda\left(1 - \sum_{x_0} \sum_{x_1} Q_1(x_0, x_1)\right).$$

Differentiating and setting equal to zero gives

$$Q_1(x_0, x_1) = \frac{1}{\lambda} \sum_{i=1}^{N} \frac{Q_1(x_0, x_1)P(R_i \mid x_1)\{x_{0i} = x_0\}}{\sum_{x_1} Q_1(x_0, x_1)P(R_i \mid x_1)},$$

$\sum \sum Q_1(x_0, x_1) = 1$ implies $\lambda = N$.

We call

$$Q_1(x_0, x_1) = \frac{1}{N} \sum_{i=1}^{N} \frac{Q_1(x_0, x_1)P(R_i \mid x_1)\{x_{0i} = x_0\}}{\sum_{x_i} Q_1(x_0, x_1)P(R_i \mid x_1)}$$

the updating equation for Q_i.

A similar calculation can be carried out for internal links. Consider the link leading into node 2. We can write

$$L_1 = \prod_{i=1}^{N} \sum_{x_1} \sum_{x_2} Q_2(x_1, x_2)P(R_i \mid x_2)P(S_i \mid x_1),$$

where

$$P(R_i \mid x_2) = P(x_{4i}, x_{5i} \mid x_2)$$
$$= P_4(x_2, x_{4i})P_5(x_2, x_{5i})$$

and

$$P(S_i \mid x_1) = P(x_{0i}, x_{6i}, x_{7i} \mid x_1)$$
$$= P(x_{0i} \mid x_1)P(x_{6i}, x_{7i} \mid x_1)$$
$$= \frac{Q_1(x_{0i}, x_1)}{\sum Q_1(x_0, x_1)}$$
$$\times \sum_{x_3} P_3(x_1, x_3)P_6(x_3, x_{6i})P_7(x_3, x_{7i}).$$

Arguing as before the updating equation for $Q_2(x_1, x_2)$ is

$$Q_2(x_1, x_2)$$
$$= \frac{1}{N} \sum_{i=1}^{N} \frac{Q_2(x_1, x_2)P(R_i \mid x_2)P(S_i \mid x_1)}{\sum_{x_1} \sum_{x_2} Q_2(x_1, x_2)P(R_i \mid x_2)P(S_i \mid x_1)}.$$

Updating equations can also be derived for $Q_3, Q_4, \cdots, Q_7$. The algorithm proceeds by starting with some initial values for all the probabilities and iterating through the sequence of updating equations until the increase in likelihood is small.

11. A GENERAL TREE

Consider a tree formed from observations on S species. Let $N = \{0, 1, 2, \cdots, M\}$ be the collection of nodes in the tree, 0 corresponding to the root. Let $\mathbf{X}_k = (x_{k1}, x_{k2}, \cdots, x_{kN})$ be the sequence of node k. We only consider trees for which $\mathbf{X}_0$ is one of the observed sequences.

The tree may be described by a function

$$\tau: N \to N$$

where $\tau(j)$ is the next node to j that is closer to the root 0. Let $E = \{j: \nexists i \in N \text{ with } \tau(i) = j\} \cdot E$ is the set of end nodes in the tree. Let P_j be the transition matrix on the link joining $\tau(j)$ and j; P_0 is the marginal probability of the root.

Assumptions 1 and 2 of Section 9 apply without change to the general tree. The likelihood of the tree may be written as

$$L = \prod_{i=1}^{N} P_0(x_{0i}) \prod_{j=1}^{M} P_j(x_{\tau(j)i}, x_{ji}).$$

In this expression all the probabilities are unknown as well as the sequences $\mathbf{x}_k$ for which $k \notin E$. As in Section 9 we can reduce the number of unknowns considerably by summing over all x_{ji} for which $j \notin E$ to get

$$L_1 = \prod_{i=1}^{N} \sum_{\{x_{ki}: k \notin E\}} P_0(x_{0i}) \prod_{j=1}^{M} P_j(x_{\tau(j)i}, x_{ji}).$$

In this expression only the probabilities are left as unknowns.

An example may help to clarify the notation. Consider the tree shown in Figure 6. Here $N = \{0, 1, 2, 3, 4, 5, 6, 7, 8\}$. $\tau(0) = 0$, $\tau(1) = 0$, $\tau(2) = 1$, $\tau(3) = 1$, $\tau(4) = \tau(5) = 2$, $\tau(6) = \tau(7) = 3$, $\tau(8) = 0$. $E = \{0, 4, 5, 6, 7, 8\}$ and

$$L = \prod_{i=1}^{N} P_0(x_{0i})P_1(x_{0i}, x_{1i})P_2(x_{1i}, x_{2i})P_3(x_{1i}, x_{3i})$$
$$\times P_4(x_{2i}, x_{4i})P_5(x_{2i}, x_{5i})P_6(x_{3i}, x_{6i})P_7(x_{3i}, x_{7i})P_8(x_{0i}, x_{8i})$$

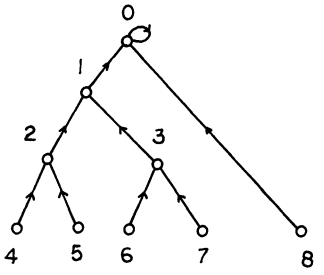

FIG. 6. *The function τ defined on 0, 1, 2, 3, 4, 5, 6, 7, 8 determines the tree. Any function τ with the property: $\tau^k i = i$ for $k \geq 1$ implies $i = 0$, will produce a tree with root 0.*

and

$$L_1 = \prod_{i=1}^{N} \sum_{x_{1i}} \sum_{x_{2i}} \sum_{x_{3i}} P_0(x_{0i})P_1(x_{0i}, x_{1i})P_2(x_{1i}, x_{2i})P_3(x_{1i}, x_{3i})$$

$$\times P_4(x_{2i}, x_{4i})P_5(x_{2i}, x_{5i})P_6(x_{3i}, x_{6i})P_7(x_{3i}, x_{7i})P_8(x_{0i}, x_{8i}).$$

Maximization of L can be carried out by applying the algorithm described in Section 10 to this case.

Maximization of L_1 is more difficult to describe. Let Q_i be the joint distribution on the link joining $\tau(j)$ and j. If we remove this link from the tree we divide the end nodes into two groups G_{1j} and G_{2j}, where G_{1j} comprises those connected to $\tau(j)$ and G_{2j} those connected to j. We need to consider three cases:

(1) $\tau(j) = 0, \quad j \neq 0.$

We can write

$$L_1 = \prod_{i=1}^{N} \sum_{x_{ji}} Q_j(x_{0i}, x_{ji})P(G_{2j} \mid x_{ji})$$

where $P(G_{2j} \mid x_{ji})$ is the conditional probability of G_{2j} given x_{ji}; here $G_{1j} = \varnothing$.

Differentiating and setting equal to zero gives

$$Q_j(x_1, x_2) = \frac{1}{N} \sum_{i=1}^{N} \frac{Q_j(x_1, x_2)P(G_{2j} \mid x_2)}{\sum_{x_2} Q_j(x_1, x_2)P(G_{2j} \mid x_2)} \{x_{0i} = x_1\}$$

as an updating equation for $Q_j(x_1, x_2)$.

(2) $j \in E \backslash \{0\}$.

Let $k = \tau(j)$. Then we can write

$$L_1 = \prod_{i=1}^{N} \sum_{x_{ki}} Q_j(x_{ki}, x_{ji})P(G_{1j} \mid x_{ki}).$$

Here $G_{2j} = \varnothing$. This leads to

$$Q_j(x_1, x_2) = \frac{1}{N} \sum_{i=1}^{N} \frac{Q_j(x_1, x_2)P(G_{1j} \mid x_1)}{\sum_{x_1} Q_j(x_1, x_2)P(G_{1j} \mid x_1)} \{x_{ji} = x_2\}$$

as an updating equation for $Q_j(x_1, x_2)$.

(3) $j \notin E, \tau(j) \notin E$.

Let $k = \tau(j)$. Then we can write

$$L_1 = \prod_{i=1}^{N} \sum_{x_{ji}} \sum_{x_{ki}} Q_j(x_{ki}, x_{ji})P(G_{1j} \mid x_{ki})P(G_{2j} \mid x_{ji}),$$

leading to

$$Q_j(x_1, x_2) = \frac{1}{N} \sum_{i=1}^{N} \frac{Q_j(x_1, x_2)P(G_{1j} \mid x_1)P(G_{2j} \mid x_2)}{\sum_{x_1} \sum_{x_2} Q_j(x_1, x_2)P(G_{1j} \mid x_1)P(G_{2j} \mid x_2)}$$

as an updating equation for $Q_j(x_1, x_2)$.

The algorithm proceeds by starting with some initial values for all the probabilities and iterating through the sequence of updating equations until the increase in likelihood is small.

When constructing trees from data on a large number of species the task of finding the optimal tree topology becomes a major computational problem. For the case of five species we can simply calculate the likelihood of each of the 15 tree topologies and choose the topology which produces the largest likelihood. This approach becomes increasingly impractical, however, as the number of species grows. Edwards and Cavalli-Sforza (1964) showed that the number of unrooted bifurcating trees with n labeled tips is $(2n - 5)!/[(n - 3)!\, 2^{n-3}]$ which for as few as 10 tips (i.e., species) is well over 2 million.

Felsenstein (1981) has described a less ambitious strategy in which the tree is built up by successively adding species to it starting with a two-species tree. When the kth species is being added to the tree there will be $2k - 5$ links from which it could arise. Each of these is tried and the maximum likelihood within the resulting topology is evaluated using either of the techniques described above. The placement yielding the highest likelihood is chosen. Before the next species is added local rearrangements are carried out in the tree to see if any of these improves the likelihood of the tree. If any does, the resulting tree is chosen and the rearrangement process continues until a tree is found which no local rearrangement can improve. This strategy is not guaranteed to find the optimal topology but is considerably shorter computationally than complete search and, according to Felsenstein (1981), works well in practice.

For our probability model in which different transition probabilities are possible on each link of the tree, this procedure will be particularly easy to apply. Consider the example in Figure 7.

Starting from the tree with nodes 0–7 suppose we add a new species (node 9) by joining it to the link between 1 and 3 at node 8.

1. *MPL.* The transition probabilities on links other than that joining 1 and 3 should be little affected by this addition so we could leave them fixed and confine attention to the subtree made up of {1, 3, 8, 9}. Then having calculated the estimates of $\mathbf{x}_8$ and the transition probabilities on links $1 \rightarrow 8$, $8 \rightarrow 3$ and $8 \rightarrow 9$ we can calculate the likelihood obtained by adding the new species at this link. Similar calculations may be carried out for the other links and the one leading to the optimal likelihood chosen. As a final step the MPL routine should be applied to the complete tree.

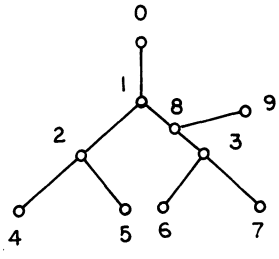

FIG. 7. *Adding a new species to the tree. Species 9 may be added to the tree within each of the seven links of the old tree.*

2. *MAL.* Before adding the new species the likelihood is

$$L_1 = \prod_{i=1}^{N} \sum_{x_{1i}} \sum_{x_{3i}} Q_{1,3}(x_{1i}, x_{3i}) P(G_1 \mid x_{1i}) P(G_2 \mid x_{3i})$$

where $Q_{1,3}$ is the joint probability on the link $1 \rightarrow 3$; $P(G_1 \mid x_{1i})$ is the conditional probability of the end nodes connected to 1 given x_{1i}; $P(G_2 \mid x_{3i})$ is similarly defined.

On adding the new species the likelihood becomes

$$L_1^1 = \prod_{i=1}^{N} \sum_{x_{1i}} \sum_{x_{3i}} \sum_{x_{8i}} Q_{1,8}(x_{1i}, x_{8i}) P_9(x_{8i}, x_{9i}) P_3(x_{8i}, x_{3i}) \times P(G_1 \mid x_{1i}) P(G_2 \mid x_{3i})$$

where P_9 is the transition probability on link $8 \rightarrow 3$. Fixing $P(G_1 \mid x_{1i})$ and $P(G_2 \mid x_{3i})$ at their values from the smaller tree, we can iterate to find $Q_{1,8}$, P_9 and P_3.

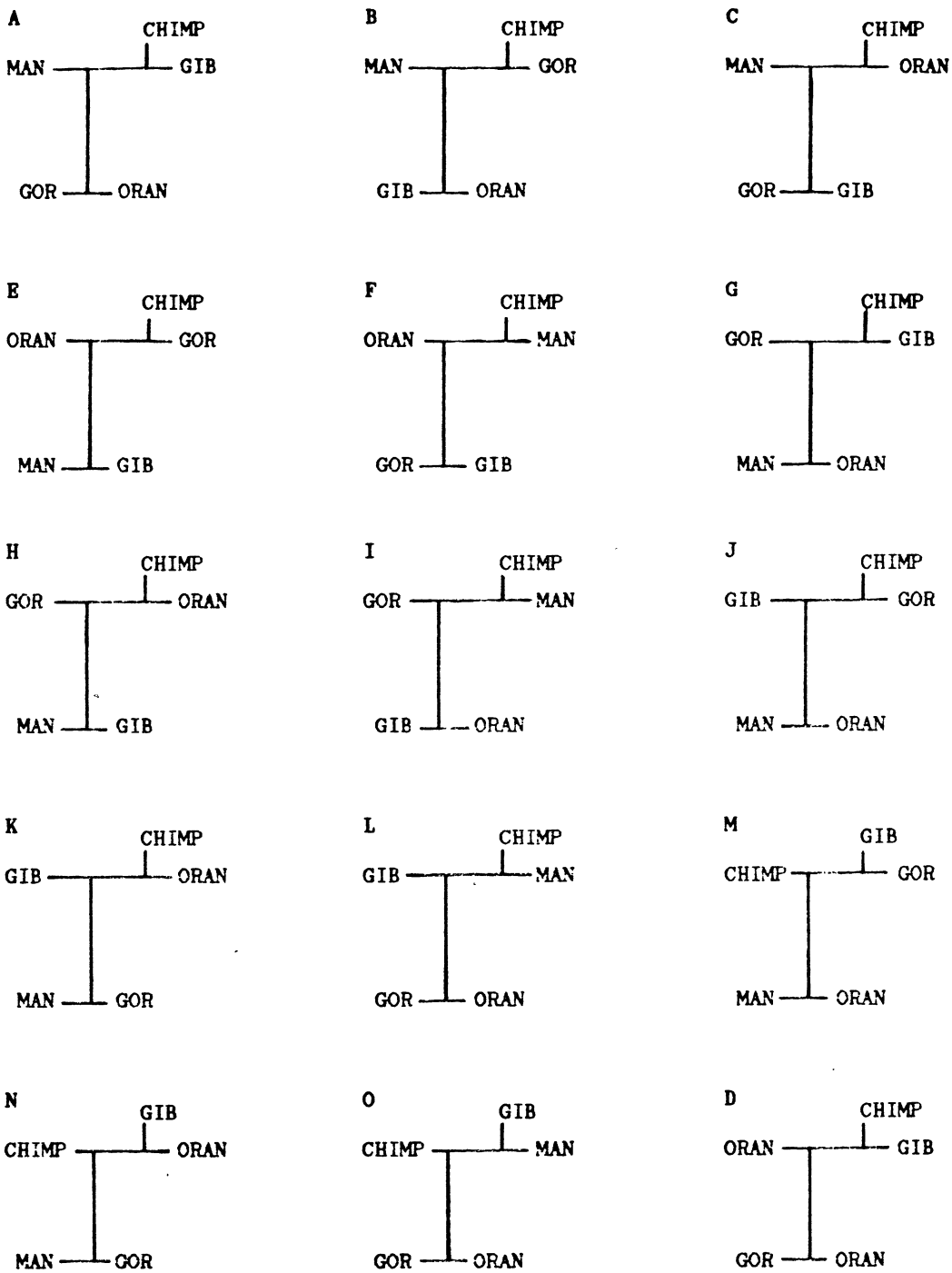

FIG. 8. *Possible evolutionary trees for five species.*

This gives a likelihood value for adding the new species to this particular link. Likelihood may be similarly calculated for the other links and the one leading to the largest likelihood chosen. As a final step the MAL routine should be applied to the full tree.

12. DATA ANALYSIS

The above techniques were applied to the sequences of mitochondrial DNA from chimpanzee, gibbon, gorilla, orangutan and human. The 15 different possible trees are shown in Figure 8 together with the labeling scheme that will be used to identify them.

Each technique was applied to the full sequence of length 896 and the results are shown in Table 13. In both cases trees I, N and B, in that order, were picked out as most likely. These are the trees for which gibbon and orangutan are together as well as two from chimpanzee, gorilla and human. Tree I (the most likely tree) has chimpanzee and man on the same branch.

In order to test the sensitivity of these results random subsamples were generated, each site having probability 0.5 of being included. The techniques were applied to each set of subsequences to calculate the likelihoods of trees I, N and B. The results are shown in Table 14. The log likelihood differences show huge variation among subsamples and the order I, N, B is maintained in only three of the five subsamples for MPL and four of the five subsamples for MAL. The reason for this instability becomes apparent when we examine the synapomorphy counts in Table 15. Six synapomorphies separating chimpanzee–human from the rest are transversions. If a good number of these transversions appear in a subsample, then the likelihood of I versus N, B will be high. If a small number appear it will be low.

Hence the separation in the likelihoods of the different trees is brought about in large part by what happens at the few sites where synapomorphies occur. If these sites are not included in a subsample, the whole picture changes.

We saw earlier that substitution rates for silent sites were much higher than those for replacement sites. It may be that the probability structure for silent sites differs from that for replacement sites. To test this we fit separate models to silent and replacement sites. The results are shown in Table 16. Based on total likelihood values the best fitting tree is again tree I. For both techniques the new model has an extra 87 parameters—3 marginal probabilities plus 7 × 12 conditional probabilities. For MPL the likelihood increase for tree I is $-2436.9 - (-2597.5) = 160.6$ and for MAL the increase is $-2411.5 - (-2571.5) = 160.0$, both of which are highly significant when compared with the values in a $\frac{1}{2}\chi^2_{87}$ table. Hence separate models for silent and replacement sites leads to a significant improvement in the fit.

One of the referees points out that the models have a rather large number of parameters; MAL has 87 parameters for the full data set and 174 for the silent and replacement sites treated separately. The data here consists of the joint distribution of nucleotides at five end nodes, making $4^5 = 1024$ counts in all; the sum of these 1024 counts is 896. Certainly this is a sparse table, since most counts occur in the four no change cells of the tables; and certainly we would not want to believe that the likelihood increase can be accurately evaluated by $\frac{1}{2}\chi^2_{87}$.

TABLE 13
Log likelihoods for 15 possible trees using full sequences ($n = 896$)

	Most parsimonious likelihood	Maximum average likelihood
A	−2653.4	−2621.0
B	−2613.8	−2582.6
C	−2658.7	−2628.7
D	−2656.7	−2622.7
E	−2663.9	−2627.4
F	−2647.0	−2611.4
G	−2648.7	−2628.6
H	−2657.3	−2633.5
I	−2597.5	−2571.5
J	−2662.5	−2635.5
K	−2660.9	−2625.5
L	−2648.0	−2611.3
M	−2653.4	−2630.2
N	−2609.0	−2580.5
O	−2655.9	−2626.3

TABLE 14
Log likelihood for trees I, N and B using random subsequences

	Tree I	Tree N	Tree B
Most parsimonious likelihood			
$n = 484$	−1382.5	−1392.5	−1393.5
$n = 458$	−1334.3	−1328.2	−1333.2
$n = 454$	−1297.7	−1298.7	−1303.3
$n = 449$	−1323.3	−1328.2	−1335.3
$n = 428$	−1195.3	−1199.7	−1198.9
Maximum average likelihood			
$n = 484$	−1369.7	−1377.4	−1378.9
$n = 458$	−1320.3	−1316.1	−1320.3
$n = 454$	−1285.4	−1287.3	−1289.1
$n = 449$	−1308.8	−1315.3	−1317.2
$n = 423$	−1182.7	−1184.0	−1187.4

TABLE 15
Synapomorphy counts

	Transitions				Transversions			
	C	Go	O	Gi	C	Go	O	Gi
H	8	7	2	3	6	0	0	0
C		11	1	5		0	0	1
Go			6	9			0	0
O				24				4

TABLE 16
Log likelihoods for silent and replacement sites for each tree

	MPL			MAL		
	Silent	Replacement	Total	Silent	Replacement	Total
A	−858.0	−1632.9	−2490.9	−841.4	−1626.1	−2467.5
B	−868.2	−1582.6	−2450.8	−839.7	−1573.8	−2413.5
C	−861.5	−1639.9	−2498.4	−840.6	−1630.5	−2471.1
D	−862.9	−1631.7	−2494.6	−839.9	−1627.2	−2467.1
E	−869.0	−1626.4	−2495.4	−841.4	−1622.6	−2464.0
F	−864.4	−1623.9	−2488.3	−835.3	−1619.9	−2455.2
G	−860.6	−1632.0	−2492.6	−843.4	−1629.3	−2472.7
H	−867.1	−1636.5	−2503.6	−841.9	−1634.6	−2476.5
I	−854.4	−1582.5	−2406.9	−834.5	−1577.0	−2411.5
J	−867.7	−1625.7	−2493.4	−843.2	−1622.1	−2465.3
K	−865.0	−1635.0	−2500.0	−840.9	−1630.6	−2471.5
L	−861.9	−1623.5	−2485.4	−835.3	−1618.9	−2454.2
M	−865.3	−1635.8	−2501.1	−843.5	−1629.3	−2472.8
N	−855.8	−1590.4	−2446.2	−836.7	−1580.7	−2417.4
O	−864.7	−1636.3	−2501.2	−840.7	−1629.7	−2470.4

For the split model the log likelihood differences among the trees I, N and B are considerably smaller than those obtained previously. Thus using MAL with the split model leads to a rather uncertain discrimination among trees I, N and B.

The results for the replacement sites are largely in agreement with those for the full sequence. However, the results for the silent sites are markedly different. Using MPL on silent sites, the range of log likelihood values obtained was $-854.4 - (-869.0) = 14.6$, while for replacement sites the range was $-1582.5 - (-1636.9) = 54.4$. Using MAL the difference was even more marked: for silent sites $-834.5 - (-843.5) = 9.0$ and for replacement sites $-1573.8 - (-1634.6) = 60.8$. Thus replacement sites serve better to discriminate among the various trees. There are so many changes in the silent sites that they don't offer much discrimination between close species.

Replacement sites favor B (chimpanzee–gorilla) over I (chimpanzee–human); silent sites do the opposite. The overall likelihoods favor chimpanzee–human slightly, but the results are still within the fuzz of statistical error.

Figure 9 shows the estimated link lengths in the MAL case for tree I. The distances for silent sites are huge due to the relatively large substitution rates at these sites. The replacement sites point to H, C and Go having a common ancestor, while for the combined data the distance between the ancestor of C and H and that of Go is only 0.01, highlighting the uncertainty about which tree is best for describing the relationship among C, H and Go. All trees show 0 and Gi well separated from C, H and Go. All trees suggest asynchronous evolution, with more evolution in the gorilla line than the human–chimpanzee line.

We need more data to decide the tree for chimpanzee, human and gorilla. It would be good to avoid the many parameters produced by a different transition matrix on every link, by plausible assumptions such as Felsenstein's (1983); but in principle, and certainly for large trees, the more general model should be one of the candidates. Finally, there is the lurking problem of dependency along the sequence; some kind of joint model predicting the value at a position based on its neighbors in the tree and along the sequence is needed.

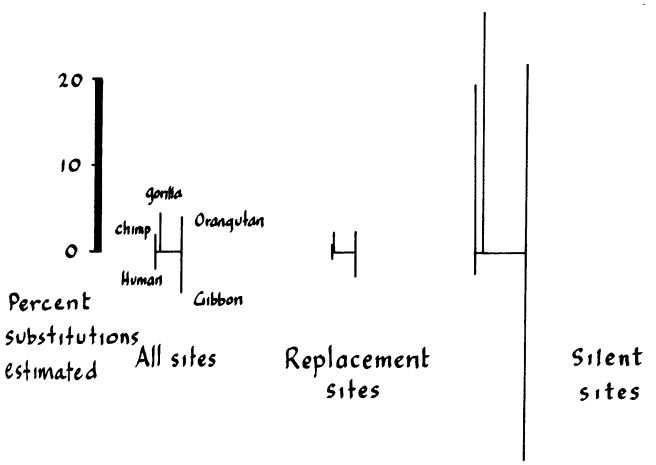

FIG. 9. *Estimated substitutions. The transition matrices within each link of the tree are estimated to maximize the average likelihood, and the proportion of substitutions in each link is estimated by* $-\frac{1}{4}$ log det *of the transition matrix. The estimated substitutions for a silent site are about ten times those for replacement sites. There is only slight evidence, mostly from silent sites, favoring the human-chimpanzee branch. In every case, asynchronous evolution is strongly suggested.*

REFERENCES

ALBERTS, B., BRAY, D., LEWIS, J., RAFF, M., ROBERTS, K. and WATSON, J. D. (1983). *Molecular Biology of the Cell.* Garland, New York.

BROWN, W. M., PRAGER, E. M., WANG, A. and WILSON, A. C. (1982). Mitochondrial DNA sequences of primates: tempo and mode of evolution. *J. Mol. Evol.* **18** 225–239.

CAVALLI-SFORZA, L. L. and EDWARDS, A. W. F. (1967). Phylogenetic analysis—models and estimation procedures. *Amer. J. Human Genet.* **19** 233–257.

DAYHOFF, M. O. (1978). Survey of new data and computer methods of analysis. In *Atlas of Protein Sequence and Structure* (M. O. Dayhoff, ed.). National Biomedical Research Foundation, Washington.

EDWARDS, A. W. F. and CAVALLI-SFORZA, L. L. (1964). Reconstruction of evolutionary trees. In *Phenetic and Phylogenetic Classification* (V. H. Heywood and J. McNeill, eds.) **6**. Systematics Association, London.

ERDÖS, P. and SZEKERES, G. (1935). A combinatorial problem in geometry. *Compositio Math.* **2** 463–470.

FELSENSTEIN, J. (1981). Evolutionary trees from DNA sequences: a maximum likelihood approach. *J. Mol. Evol.* **17** 368–376.

FELSENSTEIN, J. (1983). Statistical inference of phylogenies. *J. Roy. Statist. Soc. Ser. A* **146** 246–272.

FERRIS, S. D., WILSON, A. C. and BROWN, W. M. (1981). Evolutionary tree for apes and humans based on cleavage maps of mitochondrial DNA. *Proc. Nat. Acad. Sci. U.S.A.* **78** 2431–2436.

GOODMAN, M., ROMERO-HERRERA, A. E., DENE, H., CZELUSNIAK, J. and TASHIAN, R. E. (1982). Amino acid sequence evidence on the phylogeny of primates and other eutherians. In *Macromolecular Sequences in Systematic and Evolutionary Biology* (M. Goodman, ed.) 115–187. Plenum, New York.

HARTIGAN, J. A. (1967). Representation of similarity matrices by trees. *J. Amer. Statist. Assoc.* **62** 1140–1158.

KLUGE, A. G. (1983). Cladistics and the classification of the great apes. In *New Interpretations of Ape and Human Ancestry* (R. L. Gochon and R. S. Corruccini, eds.) 151–177. Plenum, New York.

NEYMAN, J. (1971). Molecular studies of evolution: a source of novel statistical problems. In *Statistical Decision Theory and Related Topics* (S. S. Gupta and J. Yackel, eds.) 1–27. Academic, New York.

SANKOFF, D. and KRUSKAL, J. B. (1983). *Time Warps, String Edits, and Macromolecules: The Theory and Practice of Sequence Comparison.* Addison-Wesley, London.

SIBLEY, C. G. and AHLQUIST, J. E. (1983). Phylogeny and classification of birds based on the data of DNA-DNA hybridization. In *Current Ornithology* **1** (R. F. Johnson, ed.) 245–292. Plenum, New York.

SIBLEY, C. G. and AHLQUIST, J. E. (1984). The phylogeny of the hominoid primates, as indicated by DNA-DNA hybridization. *J. Mol. Evol.* **20** 2–15.

YUNIS, J. J. and PRAKASH, O. M. (1982). The origin of man: a chromosomal pictorial legacy. *Science* **215** 1526–1530.

Comment

Stephen Portnoy

I wish to thank the authors for bringing some important statistical problems in molecular evolution to the attention of statisticians. This is an area which generates a large number of statistical modeling problems requiring a very delicate balance between sufficient complexity to explain the phenomena and sufficient simplicity to carry out statistical inference. I particularly appreciate the authors' development of Markovian models for the occurrence of specific base pairs along the DNA molecule. The notion of an "effective" sequence should have important consequences. I would suggest, however, that since effectives are most likely generated by biochemical causes, they may be constant over very wide ranges of organisms. Thus it may be possible to pool all (or very large parts of) the DNA sequence data to search for effectives. With a sufficiently large data set, it should be possible to fit arbitrary kth order Markov models (for $k = 4$ or 5) against which one could legitimately test whether or not a particular sequence is effective. Once a set of reasonably short effective sequences is found, it should be possible to build more appropriate models to analyze molecular evolution among species.

I do have a few technical quibbles about parts of the paper. First I am bothered by the use of the F distribution for analyzing the evolutionary distance measures in Section 2. It seems that the model underlying the analysis represents each distance as the sum of fixed parameters plus a (putative) iid normal error. Although the F tests possess some robustness, I believe such a model may be entirely inappropriate. Random variation occurring along each link in the tree could produce very high correlations between distances for closely related species. Clearly, the distance measures are based on data most reasonably modeled as a (Markovian) process occurring along the tree. The dependence in such a model could completely invalidate the F distribution. This type of problem was first brought to my attention by some colleagues here at the University of Illinois. A referee of a paper they had written noticed just this problem in a very closely related situation. I found the development and analysis of appropriate statistical models to be extremely interesting research (see Ferris, Portnoy and Whitt, 1979).

One other quibble is the use of χ^2 approximations in large, sparse situations. I would suggest that such results need to be justified by appropriate asymptotics (e.g., see Morris, 1975).

Stephen Portnoy is Professor of Statistics, Department of Statistics, University of Illinois at Urbana-Champaign, 725 S. Wright, Champaign, Illinois 61820.