J Mol Evol (1996) 42:587–596

JOURNAL OF MOLECULAR EVOLUTION

Maximum-Likelihood Models for Combined Analyses of Multiple Sequence Data

Ziheng Yang[1,2,*]

[1]Institute of Molecular Evolutionary Genetics and Department of Biology, The Pennsylvania State University, 328 Mueller Laboratory, University Park, PA 16802, USA
[2]College of Animal Science and Technology, Beijing Agricultural University, Beijing 100094, China

Received: 9 August 1995 / Accepted: 30 November 1995

Abstract. Models of nucleotide substitution were constructed for combined analyses of heterogeneous sequence data (such as those of multiple genes) from the same set of species. The models account for different aspects of the heterogeneity in the evolutionary process of different genes, such as differences in nucleotide frequencies, in substitution rate bias (for example, the transition/transversion rate bias), and in the extent of rate variation across sites. Model parameters were estimated by maximum likelihood and the likelihood ratio test was used to test hypotheses concerning sequence evolution, such as rate constancy among lineages (the assumption of a molecular clock) and proportionality of branch lengths for different genes. The example data from a segment of the mitochondrial genome of six hominoid species (human, common and pygmy chimpanzees, gorilla, orangutan, and siamang) were analyzed. Nucleotides at the three codon positions in the protein-coding regions and from the tRNA-coding regions were considered heterogeneous data sets. Statistical tests showed that the amount of evolution in the sequence data reflected in the estimated branch lengths can be explained by the codon-position effect and lineage effect of substitution rates. The assumption of a molecular clock could not be rejected when the data were analyzed separately or when the rate variation among sites was ignored. However, significant differences in substitution rate among lineages were found when the data sets were combined and when the rate variation among sites was accounted for in the models. Under the assumption that the orangutan and African apes diverged 13 million years ago, the combined analysis of the sequence data estimated the times for the human-chimpanzee separation and for the separation of the gorilla as 4.3 and 6.8 million years ago, respectively.

Key words: Models — Maximum likelihood — Multiple gene data — Molecular clock

Introduction

It is now commonplace for a molecular systematist to have access to multiple data sets of sequences for the species of his interest, and there is need to combine such heterogeneous data in phylogenetic analysis. One such case is when several genes are sequenced for the same species, and another case arises in analyzing a protein-coding gene in which the three codon positions have quite different evolutionary dynamics. Combined analysis can be expected to lead to more reliable estimation of phylogenetic relationships and more accurate calculation of branching dates. Data of multiple genes also provide an opportunity for testing hypotheses concerning the similarities and differences in the evolutionary dynamics of different genes. However, different genes perform different functions and may have followed different evolutionary processes. For example, nucleotide frequencies are often quite different among genes, the transition/transversion rate bias is much higher in mitochondrial

[*] *Present address:* Department of Integrative Biology, University of California, Berkeley, CA 94720-3140, USA

genes than in nuclear genes, and slowly changing conserved genes usually involve a severe among-site variation in evolutionary rates while substitution rates in pseudogenes are more or less homogeneous across sites. The heterogeneity of the substitution process among genes should be taken into account in the combined analysis.

There has been considerable controversy over whether different data sets should be analyzed separately or combined in one analysis (see, e.g., Swofford 1991; Bull et al. 1993 and references therein). While some of these discussions highlight the danger of merging heterogeneous data sets in a parsimony analysis without accommodating their heterogeneity, no generally useful approach to the problem seems to have been suggested.

In this paper, models suitable for analyzing data of multiple genes were developed. They are formulated at the level of nucleotides and account for different aspects of among-gene heterogeneity in the substitution process. Factors considered include the nucleotide frequency bias, the substitution rate bias (such as the transition/transversion rate ratio), and the difference in the extent of rate variation among sites. The maximum-likelihood framework was adopted. Although likelihood values calculated from different tree topologies under any of these models can be compared to reconstruct the phylogenetic tree (see, e.g., Felsenstein 1981), the objective of this paper is to estimate parameters and test hypotheses. Data from a segment of the mitochondrial genome from six hominoid species were analyzed. The three codon positions in the protein-coding regions and the tRNA-coding region in the segment were regarded as four heterogeneous data sets and were combined in the analysis with their heterogeneity accounted for.

Data and Methods

Data. The data consist of a segment of the mitochondrial DNA (mtDNA) from human, common chimpanzee, pygmy chimpanzee, gorilla, orangutan, and siamang, published by Horai et al. (1992) and prepared by Takezaki (Takezaki et al. 1995). The phylogenetic relationship among the species appears to be well established, and the tree shown in Fig. 1 will be assumed in the analysis. The sequence data contain the complete cDNAs for NADH dehydrogenase subunit 2 (*ND2*), cytochrome oxidase subunits I and II (*COI* and *COII*), *ATPase 8,* portions of two genes for *ND1* and *ATPase 6,* and 11 interspersed tRNA genes. The *ATPase 8* and *ATPase 6* genes overlap for 43 nucleotides, using different reading frames. These nucleotides were used twice in the two genes in this paper, whereas Horai et al. (1992) deleted them. After exclusion of alignment gaps, each sequence contained 1,367 codons and 739 nucleotides from the tRNA coding regions (Table 1).

Codon and base frequencies were quite homogeneous across species and among the protein-coding genes. In this paper, differences among the six genes were ignored, and nucleotides at the three codon positions and those from the tRNA-coding regions were considered four different data sets, to be combined in one analysis with their heterogeneity taken into account. The four site classes will be loosely referred to as ‘‘codon positions.’’

Models. The theory will be described in the context of combining data of multiple genes, although in the case of the example data set, different codon positions rather than different genes are combined. So the terms ‘‘gene’’ and ‘‘codon position’’ are used interchangeably in the rest of this paper. The unrooted tree of Fig. 1A will be used as an example, where $\mathbf{t} = \{t_1, t_2, t_3, t_4, t_5, t_6, t_7, t_8, t_9\}'$ represents the branch lengths. The substitution model of Hasegawa et al. (1985) will be assumed, by which the substitution rate from nucleotide i to j ($i \neq j$) is given by

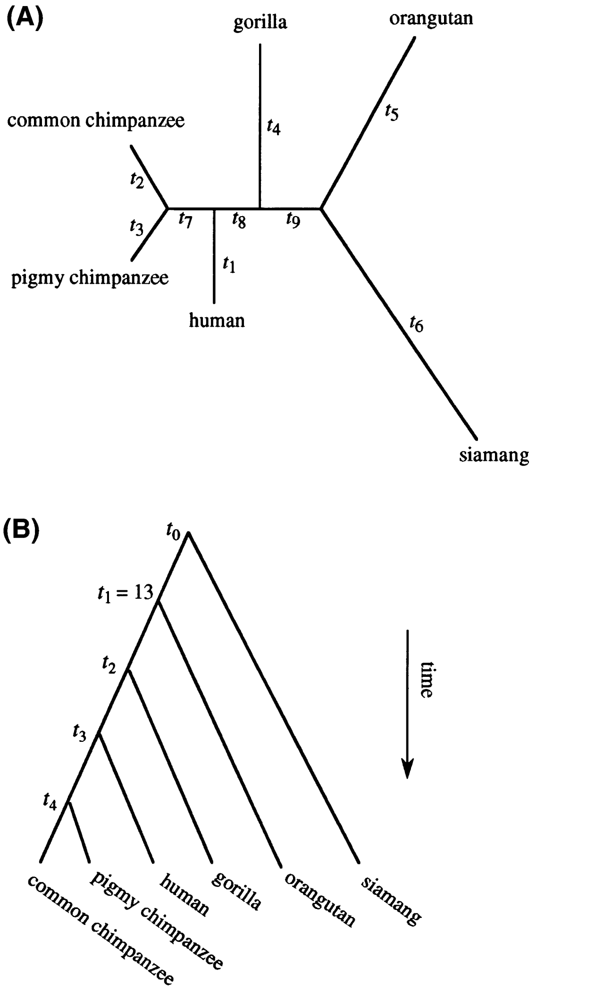

Fig. 1. Phylogenetic trees for six hominoid species, whose sequence data were analyzed in this paper. **A** The unrooted topology, used when the molecular clock (rate constancy among lineages) was not assumed in the models. Substitution rates are allowed to vary among lineages and the nine branch lengths are free parameters. **B** The rooted tree, used when the molecular clock was assumed in the models. One evolutionary rate is assumed throughout the tree, and the parameters are the five divergence times. Branch lengths **(A)** or divergence times **(B)** are measured by the average numbers of nucleotide substitutions per site.

$$Q_{ij} = \begin{cases} \kappa\pi_j & \text{(for transitions: } T \leftrightarrow C, A \leftrightarrow G) \\ \pi_j & \text{(for transversions: } T \leftrightarrow A, T \leftrightarrow G, C \leftrightarrow A, C \leftrightarrow G) \end{cases} \tag{1}$$

where j is the equilibrium frequency of nucleotide j. Parameter κ (equivalent to α/β in the notation of Hasegawa et al. 1985) is the transition/transversion rate ratio. The matrix is multiplied by a constant so that the average number of substitutions is one when the process is in equilibrium, and then time or the branch length in a tree is measured by the average number of nucleotide substitutions per site.

Table 1. Number of nucleotides, number of site patterns, and nucleotide frequencies at the three codon positions and in the tRNA genes for the hominoid mtDNAs

			Frequencies			
	# sites	# patterns	*T*	*C*	*A*	*G*
Position 1	1367	91	0.1976	0.2699	0.3198	0.2126
Position 2	1367	53	0.3994	0.2876	0.1980	0.1150
Position 3	1367	203	0.1807	0.3998	0.3686	0.0510
tRNA region	759	62	0.2620	0.2468	0.3377	0.1535
Total	4860	241	0.2597	0.3078	0.3021	0.1305

The most-general model for a combined analysis of data from multiple genes uses independent parameters for each gene. This model involves, for each gene g, an independent set of branch lengths $\mathbf{t}^{(g)} = \{t_1^{(g)}, t_2^{(g)}, t_3^{(g)}, t_4^{(g)}, t_5^{(g)}, t_6^{(g)}, t_7^{(g)}, t_8^{(g)}, t_9^{(g)}\}'$, an independent set of nucleotide frequency parameters $\pi = \{ \pi_T^{(g)}, \pi_C^{(g)}, \pi_A^{(g)}, \pi_G^{(g)}\}'$ (with the restriction that the sum is one, and a transition/transversion rate ratio parameter (κ). Let the data for gene g be $D^{(g)}$, and then the log-likelihood function is

$$\ell = \log\{\prod_g \text{prob}(D^{(g)}|\mathbf{t}^{(g)},\pi^{(g)},\kappa^{(g)}\}$$
$$= \sum_g \log\{\text{prob}(\text{D}^{(\text{g})}|\mathbf{t}^{(g)},\pi^{(g)},\kappa^{(g)})\}$$

This model is in effect a separate analysis, with the same model fitted to and parameters estimated from each of the data sets independently. At the other extreme, all parameters can be assumed to be equal for different genes, that is, $\mathbf{t}^{(g)} = \mathbf{t} = \{t_1, t_2, t_3, t_4, t_5, t_6, t_7, t_8, t_9\}'$, $\pi^{(g)} = \pi = \{\pi_T, \pi_C, \pi_A, \pi_G\}'$, and $\kappa^{(g)} = \kappa$. This model then assumes complete heterogeneity of the evolutionary process among genes. Between these two extremes lie a hierarchy of models that assume some aspects of the evolutionary process are the same while other aspects are different (Table 2). For example, model 2 assumes different nucleotide frequencies but the same transition/transversion rate ratio. Models 1, 2, and 3 all assume that branch lengths for different genes are proportional, that is, $\mathbf{t}^{(g)} = c_g\mathbf{t}$, so that elements of $\mathbf{t}$ represent *lineage-specific* substitution rates, and c_g a *gene-specific* rate. To avoid use of too many parameters, $c_1 = 1$ is fixed and other cs are rate ratios.

Models described above can be combined with the gamma distribution of substitution rates among sites in a likelihood analysis (Yang 1993). In this paper, the discrete-gamma model of Yang (1994c) is assumed, with eight rate categories used. The distribution involves a shape parameter α, which is inversely related to the extent of rate variation among sites (Wakeley 1993; Yang 1993). When both the gene-specific rate of parameters (c_gs) and the gamma rates for sites are assumed, the former accounts for the *among-gene* rate difference while the latter accounts for the *within-gene* rate variation. It is known that in conserved genes (or at the first and second codon positions) substitution rates are extremely variable at sites, while in pseudogenes (or at the third codon positions) substitution rates are more homogeneous. When data of this nature are combined in one analysis, it may be necessary to use a separate gamma distribution for each gene (model 3′).

It may be worthwhile to mention that the idea of combining heterogeneous data while accounting for their differences employed in this paper is the same as that used in the analysis of variance or multiple linear regression in elementary statistics. The sum of squares (which represents the observed variation in the dependent variable) in the analysis of variance corresponds to the log likelihood in the present models, and the F-statistics for testing the significance of factors correspond to the χ^2-distributed statistics for the likelihood ratio test. While models used in the analysis of variance are linear and the effects of factors are additive (e.g., Searle 1971), models of this paper are highly nonlinear, and the effects of factors on substitution rates are multiplicative. For example, the substitution rate from nucleotide i to j at a site from gene g with gamma rate r is $c_g r Q_{ij}$.

Calculation of the likelihood function under the models described above can be adapted from previous likelihood procedures, i.e., Felsenstein (1981) for models assuming a single rate for all sites and Yang (1994c) for models assuming (discrete-) gamma rates at sites. The process of nucleotide substitution is assumed to be homogeneous (over time) and at equilibrium, and the nucleotide frequency parameters were estimated using the observed frequencies shown in Table 1. Other parameters were estimated by numerical maximization of the likelihood function.

Results

Separate Analyses of Data from Different Codon Positions

The separate analysis serves to reveal similarities and differences in the evolutionary processes of different data sets and provides guidelines for combining data sets while accommodating their heterogeneity. The model of Hasegawa et al. (1985) was fitted to each data set by the method of maximum likelihood, either with or without the assumption of a molecular clock and either assuming a single rate for all sites or gamma rates among sites. The results are shown in Table 3. When a constant rate was assumed for all sites and when the molecular clock was not assumed, estimates of κ were 11.267 ± 1.772, 9.272 ± 2.375, 29.275 ± 2.819, and 27.575 ± 7.991 for codon positions 1, 2, and 3, and the tRNA-coding region, respectively. The differences in the transition/transversion rate bias at codon positions are clearly due to selectional constraints at the amino acid level, and most likely, the estimate for the third codon position reflects more accurately the mutational bias than those at the first and second positions. The tree length, calculated by summing the estimated branch lengths along the tree, measures the amount of evolution at the position. The estimates are in the proportion 1 : 0.387 : 8.631 : 1.004, suggesting that the second position changed at a rate about one-third that at the first position, and the third position evolved over 20 times faster than the second position.

Using the gamma distribution to accommodate the rate variation among sites greatly improved the fit of the model for all codon positions (Table 3b). Even at the

Table 2. Models and their parameters for combined analysis of multiple gene data

		Parameters			
Model	Descriptions	Branch lengths	Frequency parameters	Transition/ transversion rate ratio	If gamma rates
0	Complete homogeneity (mixed data)	$\mathbf{t}^{(g)} = \mathbf{t}$	$\pi^{(g)} = \pi$	$\kappa^{(g)} = \kappa$	α
1	Different rate parameters	$\mathbf{t}^{(g)} = c_g\mathbf{t}$	$\pi^{(g)} = \pi$	$\kappa^{(g)} = \kappa$	α
2	Different rates and frequencies	$\mathbf{t}^{(g)} = c_g\mathbf{t}$	$\pi^{(g)}$	$\kappa^{(g)} = \kappa$	α
3	Different rates, frequencies, and transition/transversion rate bias (proportional branch lengths)	$\mathbf{t}^{(g)} = c_g\mathbf{t}$	$\pi^{(g)}$	$\kappa^{(g)}$	α
3′	Proportional branch lengths and heterogeneous gamma	$\mathbf{t}^{(g)} = c_g\mathbf{t}$	$\pi^{(g)}$	$\kappa^{(g)}$	$\alpha^{(g)}$
4	Complete heterogeneity (separate analysis)	$\mathbf{t}^{(g)}$	$\pi^{(g)}$	$\kappa^{(g)}$	$\alpha^{(g)}$

Table 3. Log-likelihood values and parameter estimates obtained from separate analyses of data at codon positions

(a) Single rate among sites

	Without clock			With clock		
	ℓ	$\hat{\kappa}$	$\hat{S}$	ℓ	$\hat{\kappa}$	$\hat{S}$
Position 1	−3,355.23	11.267	0.238	−3,357.32	11.275	0.238
Position 2	−2,459.02	9.272	0.092	−2,460.97	9.245	0.092
Position 3	−5,637.98	29.275	2.054	−5,641.07	29.047	2.036
tRNA region	−1,794.49	27.575	0.239	−1,796.12	27.398	0.237

(b) Gamma rates among sites

	Without clock				With clock			
	ℓ	$\hat{\kappa}$	$\hat{\alpha}$	$\hat{S}$	ℓ	$\hat{\kappa}$	$\hat{\alpha}$	$\hat{S}$
Position 1	−3,305.36	17.641	0.174	0.368	−3,307.48	17.574	0.176	0.363
Position 2	−2,435.90	11.875	0.073	0.121	−2,439.05	11.289	0.082	0.115
Position 3	−5,621.25	52.511	1.606	3.701	−5,623.77	52.459	1.578	3.688
tRNA region	−1,769.38	45.129	0.184	0.385	−1,773.08	40.112	0.212	0.345

third codon position, substitution rates are significantly variable among sites ($2\Delta\ell = 33.46$ compared with $\chi^2(1\%) = 6.64$ with one degree of freedom). Estimates of α were 0.174 ± 0.034, 0.073 ± 0.033, 1.606 ± 0.352, and 0.184 ± 0.046 for the first, second, and third codon positions, and the tRNA region, respectively. Substitution rates are most variable at the second position and least variable at the third position. Calculation of the tree length under the gamma model suggests that, on average, 0.121 substitutions per site have occurred at the second codon position, while this number is 3.701 at the third position. These estimates are much larger than those obtained under the assumption of rate constancy at sites (Table 3a). Estimates of κ under the gamma-rates model are also larger than those obtained under the constant-rate model, especially for the third codon positions and sites in the tRNA region, where the transition/transversion rate bias is large. These results agree with previous findings that ignoring the rate variation among sites causes underestimation of branch lengths and the transition/transversion rate ratio (e.g., Wakeley 1994; Yang et al. 1994, 1995).

Estimates of parameters (such as κ, S, and α) obtained under the assumption of the molecular clock are similar to those obtained without the clock assumption (Table 3a and b). Test of the molecular clock and estimation of divergence times will be discussed in a later section.

In conclusion, the nucleotide frequencies, the transition/transversion rate bias, the amount of evolution reflected in branch lengths, and the gamma parameter which measures the extent of rate variation among sites are quite different at different codon positions. The substitutional processes must be very different at the three codon positions and at the tRNA-coding region.

Combined Analysis of Data from Different Codon Positions

Estimation of Parameters

Maximum likelihood estimates of parameters under various models are listed in Table 4 for combined analyses of the data sets. Results of Table 4a obtained under the assumption of rate constancy among sites and without the assumption of the molecular clock will be described first, and results obtained under other models will

Table 4. Log-likelihood values and parameter estimates obtained under different models in combined analyses of the mtDNA data[a]

	#p	ℓ	$\hat{c}_2$	$\hat{c}_3$	$\hat{c}_4$	$\hat{\kappa}_1$	$\hat{\kappa}_2$	$\hat{\kappa}_3$	$\hat{\kappa}_4$	$\hat{\alpha}$
					Single rate for sites					
					(a) Without clock					
Model 0	13	−14,580.66	1	1	1			12.062		
Model 1	16	−13,605.91	0.358	6.605	0.992			16.232		
Model 2	25	−13,277.25	0.384	7.746	0.985			22.877		
Model 3	28	−13,260.09	0.385	8.337	0.995	11.258	9.279	28.495	27.459	
Model 4	52	−13,246.72	0.387	8.630	1.004	11.267	9.272	29.275	27.575	
					(b) With clock					
Model 0	9	−14,582.79	1	1	1			12.072		
Model 1	12	−13,609.23	0.358	6.559	0.992			16.132		
Model 2	21	−13,283.12	0.384	7.654	0.985			22.629		
Model 3	24	−13,266.49	0.386	8.195	0.995	11.286	9.259	27.969	27.402	
Model 4	36	−13,255.48	0.386	8.547	0.995	11.275	9.245	29.047	27.398	
					Gamma rates for sites					
					(c) Without clock					
Model 0	14	−14,259.52	1	1	1			23.529		0.230
Model 1	17	−13,558.35	0.333	9.371	0.977			23.378		1.043
Model 2	26	−13,240.91	0.366	10.945	0.977			30.094		1.407
Model 3	29	−13,203.53	0.361	16.935	0.998	12.463	9.638	63.809	30.533	0.931
Model 3′	32	−13,145.15	0.312	9.240	0.962	17.895	11.579	49.774	42.231	
Model 4	56	−13,131.89	0.328	10.057	1.046	17.641	11.875	52.511	45.121	
					(d) With clock					
Model 0	10	−14,264.94	1	1	1			22.677		0.239
Model 1	13	−13,563.85	0.335	9.116	0.979			22.744		1.085
Model 2	22	−13,246.84	0.366	10.782	0.977			32.621		1.417
Model 3	25	−13,210.63	0.362	16.297	0.999	12.471	9.583	61.229	30.330	0.952
Model 3′	28	−13,153.18	0.313	9.051	0.972	17.902	11.416	47.944	41.744	
Model 4	40	−13,143.38	0.317	10.159	0.950	17.574	11.289	52.459	40.112	

[a] #p is the number of parameters in the model. Estimates of α for the four site classes under model 3′ are 0.168, 0.079, 1.707, and 0.197 when the molecular clock is not assumed (c), and are 0.171, 0.081, 1.739, and 0.198 when the clock is assumed (d). Branch lengths and divergence times are not shown. In model 4, rates for codon positions are not parameters in the model, and the values shown here are calculated as the ratios of the tree lengths estimated in the separate analysis (Table 3); for example, $\hat{S}_2/\hat{S}_1$ is listed for $\hat{c}_2$

be discussed in comparison with them. When complete homogeneity was assumed for the combined sequence data (model 0), the estimate of κ is 12.062. This falls within the range of estimates of the parameter obtained in the separate analyses (Table 3a), but is much smaller than the average of the separate estimates. Clearly, mixing data at different codon positions tends to make the method overlook some of the transitional substitutions at the third codon position and in the tRNA region where the rate bias is very high. Use of rate parameters for codon positions alleviated this problem to some extent ($\hat{\kappa} = 16.232$ by model 1 of Table 4a), and the estimate became even greater when nucleotide frequency differences at codon positions were also taken into account ($\hat{\kappa} = 22.877$ by model 2). These results parallel the previous observation that ignoring the nucleotide frequency bias led to underestimation of the transition/transversion rate bias (Yang et al. 1995). In model 3 (Table 4a), different rate parameters, nucleotide frequencies, and transition/transversion rate ratios were used for different codon positions, and the only constraint was the proportionality of branch lengths at codon positions (see Table 2). Estimates of the four κ parameters under this model were very similar to those obtained in the separate analysis (Table 3a, also shown as estimates for model 4 in Table 4a). Substitution rates at the codon positions were estimated to be in the proportion 1 : 0.385 : 8.337 : 0.995, which is very close to the ratios of the tree length calculated for different positions in the separate analysis (i.e., $\hat{S}_1 : \hat{S}_2 : \hat{S}_3 : \hat{S}_4 = 1 : 0.387 : 8.631 : 1.004$).

Results listed in Table 4c for models 0–3 were obtained under the gamma model of rates at sites. In models 1–3 where rate parameters (c_gs) are used for codon positions, the gamma distribution accounts for the remaining among-site variation after the correction for the effect of codon positions. Adding the gamma distribution affected parameter estimates considerably. Estimates of κ all became larger, as found in the separate analyses. Estimates of rate parameters for codon positions under the gamma models were similar to, but more extreme than, those of Table 4a where a single rate was assumed for all sites at the same codon position. As the separate analyses clearly suggested different α parameters for codon positions, use of one gamma distribution in model 3 of Table 4c was unrealistic and caused distorted estimates of the transition/transversion rate ratios. The estimate (0.931) of α is too large for the first and second codon positions and the tRNA genes and is too small for the third codon position. The negative correlation between estimates of κ and α, reported by Yang et al. (1994), caused κ_1, κ_2, and κ_4 to be underestimated and κ_3 to be overestimated. Similarly, estimates of the rate parameters for codon positions under this model were also unreliable. Model $3'$ uses a different gamma distribution for each position and estimates of the rate parameters and transition/transversion rate ratios at different codon positions by this model were quite similar to those obtained in the separate analysis (Table 3b or model 4 of Table 4c). Estimates of the four gamma parameters were 0.168 ± 0.029, 0.079 ± 0.033, 1.707 ± 0.337, and 0.197 ± 0.048, which are also very close to those obtained in the separate analysis (Table 3b).

Similar to the separate analyses, parameter estimates obtained under the assumption of a molecular clock (Table 4b and d) were very similar to those obtained under corresponding models without assuming the clock (Table 4a and c). Test of the molecular clock and estimation of divergence times will be discussed in a later section.

Comparison of Models

The likelihood ratio test can be used to compare different models. If the full model contains p parameters and its log likelihood is ℓ_1, and the submodel which makes q restrictions about the parameters in the full model has log likelihood ℓ_0, then twice the log-likelihood difference, known as the deviance $D = 2\Delta\ell = 2(\ell_1 - \ell_0)$, can be compared with the χ^2 distribution with $p - q$ degrees of freedom to test whether the full model fits the data significantly better than the submodel. A simpler model provides a more-parsimonious explanation of the data and is preferred if its fit to data is not significantly poorer than a more-complex model. A hierarchy of models, such as those of Table 4, can be compared by the likelihood ratio test, and the method, very similar to analysis of variance, is known as *analysis of deviance* (McCullagh and Nelder 1989). For example, adding rate parameters for codon positions (model 1 of Table 4a) increases the number of parameters by three and increases the log-likelihood by $\Delta\ell = 974.75$. The deviance $D = 2\Delta\ell = 1{,}949.50$ can be compared with a critical value of the χ^2 distribution with $df = 3$ to test whether model 1 fits the data significantly better than model 0. The tremendous improvement in the log-likelihood suggests that substitution rates at the codon positions are quite different. Comparisons of other models suggest that both nucleotide frequencies (e.g., comparison between models 1 and 2 of Table 4a) and the transition/transversion rate ratios (e.g., comparison between models 2 and 3 of Table 4a) are significantly different at the codon positions.

Table 5. Test of the assumption that branch lengths for different codon positions are proportional[a]

	$2\Delta\ell$	d.f.	χ^2 (1%)
(a) Without clock, single rate for sites	26.74	24	42.98
(b) With clock, single rate for sites	22.02	12	26.22
(c) Without clock, gamma rates for sites	26.22	24	42.98
(d) With clock, gamma rates for sites	19.60	12	26.22

[a] Models 3 and 4 (Table 4a,b) were compared when a single rate is assumed for all sites, while models $3'$ and 4 (Table 4c,d) were compared when gamma rates are assumed for sites. All tests are insignificant

Comparison of models of Table 4a with those of Table 4c (or between Table 4b and d) is a test of rate constancy among sites. As in the separate analyses, these tests are all significant, with very large likelihood differences. There is no doubt that substitution rates are variable among sites. Model $3'$ assumes different gamma distributions for codon positions and fits the data significantly better than model 3, which uses one gamma distribution across the codon positions [comparison between $2\Delta\ell = 116.76$ with $\chi^2(1\%) = 11.35$ with $df = 3$]. The extent of rate variation among sites is not the same among codon positions, as estimates of α in the separate analyses have suggested.

Table 5 shows results for testing the assumption that branch lengths for different codon positions are proportional. The full model in comparison is model 4 of Table 2 (separate analysis) and the submodel is model 3 (or $3'$), which involves the only constraint that branch lengths are proportional, i.e., $\mathbf{t}^{(g)} = c^{(g)}\mathbf{t}$. Different nucleotide frequencies, transition/transversion rate ratios, and gamma parameters in the case of gamma rates for sites are assumed in both models. The assumption of proportional branch lengths cannot be rejected whether or not a molecular clock is assumed, and whether a constant rate is assumed for all sites or gamma rates are assumed for sites (Table 5). In other words, codon-position effect and lineage effect can explain the amount of evolution that has occurred at different codon positions along different branches of the tree, despite the fact that substitution processes are quite different at different codon positions. The proportionality of branch lengths allows data from different codon positions to be combined for calculating divergence times.

Test of Molecular Clock and Estimation of Divergence Times

Log-likelihood differences for testing the assumption of the existence of a molecular clock are shown in Table 6. The deviance $(2\Delta\ell)$ due to relaxing the clock assumption can be compared with a χ^2 distribution of four degrees of freedom to test whether the clock assumption is acceptable. The clock assumption cannot be rejected when the

Table 6. Test of the molecular clock and estimation of divergence times under different models[a]

Data	$2\Delta\ell$	$\hat{t}_0$	$\hat{t}_2$	$\hat{t}_3$	$\hat{t}_4$
(a) Separate analysis of data from different codon positions					
Single rate for sites					
position 1	4.18	13	6.602 ± 0.641	3.563 ± 0.475	1.565 ± 0.341
position 2	3.90	13	5.585 ± 0.953	3.576 ± 0.749	1.422 ± 0.517
position 3	6.18	18.059 ± 1.623	7.421 ± 0.446	5.162 ± 0.333	1.648 ± 0.158
tRNA regions	3.26	15.770 ± 1.549	7.726 ± 1.030	5.041 ± 0.857	2.346 ± 0.639
Gamma rates for sites					
position 1	4.24	13	4.923 ± 0.657	2.557 ± 0.406	1.066 ± 0.252
position 2	6.30	13	4.465 ± 0.900	3.043 ± 0.717	1.128 ± 0.427
position 3	5.04	19.254 ± 3.584	6.224 ± 0.836	3.790 ± 0.416	0.990 ± 0.106
tRNA regions	7.40	18.905 ± 4.077	6.681 ± 1.193	4.413 ± 0.925	1.847 ± 0.558
(b) Combined analysis					
Single rate for sites					
Model 0	4.26	13.801 ± 0.422	8.119 ± 0.303	5.781 ± 0.260	2.293 ± 0.174
Model 1	6.64	14.394 ± 0.894	7.554 ± 0.516	5.137 ± 0.380	1.936 ± 0.190
Model 2	11.74	14.336 ± 0.898	6.980 ± 0.495	4.570 ± 0.353	1.628 ± 0.166
Model 3	12.80	14.443 ± 0.905	6.829 ± 0.493	4.417 ± 0.350	1.588 ± 0.163
Gamma rates for sites					
Model 0	10.85	15.925 ± 1.420	6.497 ± 0.449	4.280 ± 0.293	1.419 ± 0.126
Model 1	11.00	14.850 ± 1.084	6.765 ± 0.527	4.286 ± 0.368	1.431 ± 0.160
Model 2	11.86	14.661 ± 1.035	6.442 ± 0.503	3.934 ± 0.342	1.230 ± 0.141
Model 3	14.20	14.927 ± 1.098	6.022 ± 0.501	3.404 ± 0.324	0.924 ± 0.125
Model 3′	16.05	16.018 ± 2.020	5.638 ± 0.640	3.420 ± 0.414	1.033 ± 0.150

[a] The divergence time for orangutan is assumed to be $\hat{t}_1 = 13$ million years ago, and other times in the tree of Fig. 1B were estimated under the assumption of a molecular clock. In separate analyses of data from the first and second codon positions, the tree topology of Fig. 1B converges to a trifurcating tree with $\hat{t}_0 = \hat{t}_1$. χ^2 (1%) = 13.28 with $df = 4$

data are analyzed separately, whether or not the among-site rate variation is taken into account in the models (Tale 6a). For combined analyses, the test is insignificant under all models when the rate variation among sites is ignored, but is significant for models 3 and 3′, when the gamma distribution(s) is assumed for rates among sites (Table 6b). Branch lengths obtained without the clock assumption are shown in Fig. 2 and indicate higher evolutionary rates in the orangutan and siamang lineages.

Estimates of divergence times in the tree of Fig. 1B obtained under the clock assumption are also shown in Table 6. The time for the orangutan separation is fixed at 13 million years ago (Pilbeam 1986). In the separate analyses of the first and second codon positions under the clock assumption, the tree of Fig. 1B converges to a trifurcating tree with estimates of t_0 and t_1 approaching the same value, whether or not the among-site rate variation is taken into account. The maximum likelihood tree for both codon positions is different from that of Fig. 1B and has orangutan and siamang clustered as a sister clade. Calculation of the divergence times from these two positions may not be reliable.

Estimates of divergence times obtained from different codon positions are different but they involve large sampling errors. More striking are the differences of estimates obtained under different models. It is well known that simple and unrealistic models tend to underestimate sequence distances and branch lengths, and the underestimation is more serious for long branches than for short ones. In particular, ignoring the variation of substitution rates among sites has this effect (e.g., Tateno et al. 1994; Yang et al. 1994). Thus simple and unrealistic models are expected to overestimate divergence times younger than the reference time (i.e., t_2, t_3, t_4 in the tree of Fig. 1B) and to underestimate divergence times older than the reference time (t_0 in the tree of Fig. 1B) (see also Adachi and Hasegawa 1995). This pattern is apparent in Table 6, especially when estimates obtained with and without assuming gamma rates at sites are compared.

If the violation of the clock assumption is ignored, estimates of divergence times from model 3′ (the simplest model not rejected) were $\hat{t}_0 = 16.02 \pm 2.02$ for the siamang separation, $\hat{t}_2 = 5.64 \pm 0.64$ for the gorilla separation, $\hat{t}_3 = 3.42 \pm 0.41$ for the separation of the human from the chimpanzee, and $\hat{t}_4 = 1.03 \pm 0.15$ for the separation of the two chimpanzee species. The estimated times for the human-chimpanzee separation and for the gorilla separation are very close to the estimates (3.60 ± 0.58 and 5.83 ± 0.72, respectively) obtained by Adachi and Hasegawa (1995) in their maximum-likelihood analysis of the amino acid sequences in the same segment of the mtDNA. Estimates of these two

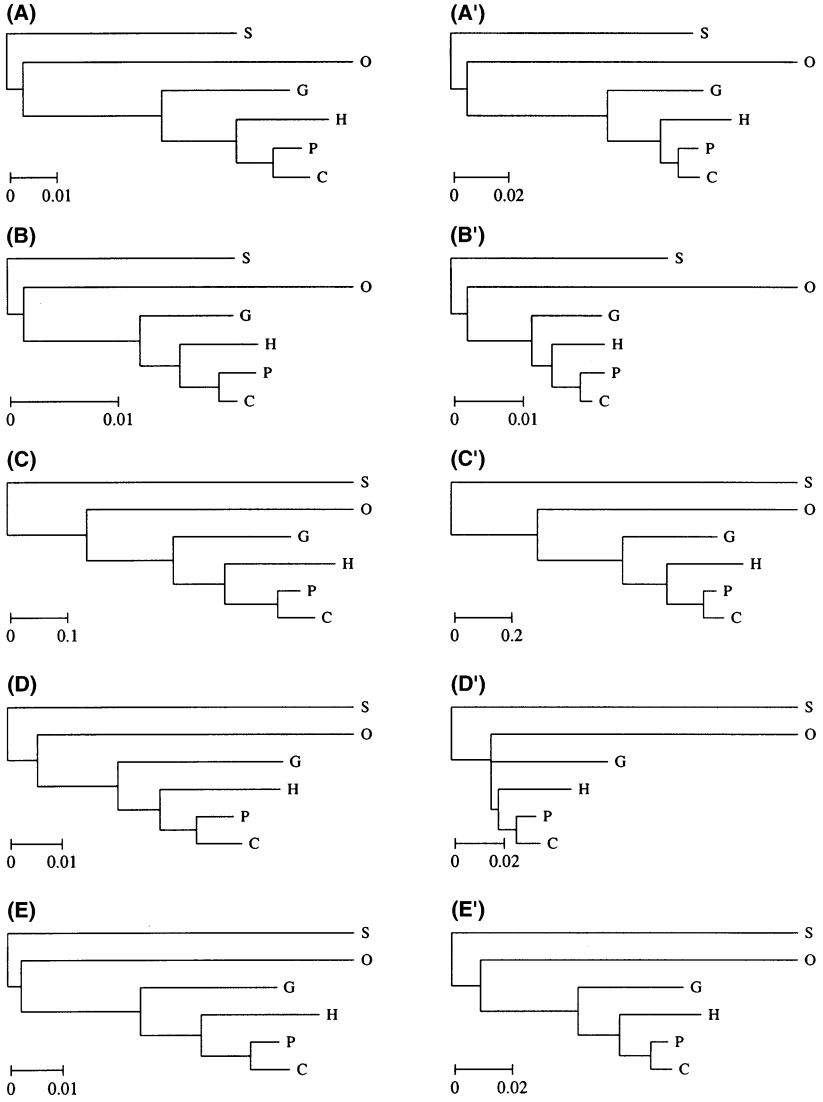

Fig. 2. Estimates of branch lengths without assuming the molecular clock (rate constancy among lineages). **A, B, C,** and **D** are from separate analysis of the first, second, and third codon positions and the tRNA-coding genes, respectively, while **E** is from the combined analysis. **A**′, **B**′, **C**′, **D**′, and **E**′ are from models assuming the gamma distribution of rates at sites.

divergence times obtained by Horai et al. (1992) (4.7 ± 0.5 and 7.7 ± 0.7) and by Takezaki et al. (1995) (5.12 ± 0.53 and 7.85 ± 0.64) are larger than the estimates of this paper. Both studies analyzed the same data and used methods based on pairwise distance estimates. The differences among the estimates seem to have two causes. First, Horai et al. (1992) and Takezaki et al. (1995) failed to properly account for the rate variation among sites, and so their estimates are expected to be overestimates. Second, estimates for model 3′ of Table 6 were obtained assuming the molecular clock, but the orangutan lineage seems to have an accelerated substitution rate (Fig. 2), and so the two divergence times appear to be underestimated under model 3′. A model that assumes rate constancy among lineages except for an independent rate for the orangutan lineage was thus fitted to the branch lengths estimated under model 3′ without assuming the clock by using a least-squares criterion, and the estimates were $\hat{t}_0 = 18.41$, $\hat{t}_2 = 6.80$, $\hat{t}_3 = 4.29$, and $\hat{t}_4 = 1.58$.

Discussions

The Reliability of the χ^2 Approximation to the Likelihood Ratio Test

Under most models used in phylogenetic analysis, data at sites are assumed to follow a multinomial distribution. The categories in the distribution correspond to the 4^s possible site patterns, where *s* is the number of species. This number is often larger than the number of sites in the sequence. Furthermore, for typical DNA sequence data, most sites have the ‘‘constant’’ patterns occupied by identical nucleotides in all species. There are then many categories with no or very few data points assigned for them. This sparseness of the data appears to have a drastic effect on the χ^2 approximation to the likelihood ratio statistic for testing the goodness of fit of a model, tending to reject the model much too often (Reeves 1992; Goldman 1993). For comparison of two closely related parametric models, Goldman (1992) conjectured that the χ^2 approximation may also be affected by the sparseness of the data. This was later noted not to be the case, as the theoretical distribution of the test statistic obtained by Monte Carlo simulation of Goldman (1993) appeared to be closely matched by the appropriate χ^2 distribution (Yang et al. 1995). In fact, the reliability of the χ^2 approximation to the likelihood ratio statistic for comparing two nested parametric models in case of sparse data is well established in statistics (Haberman 1977; Agresti and Yang 1987). Tests considered in this paper are all of this sort.

The χ^2 approximation, however, may not be reliable when the submodel is equivalent to fixing some parameters at the boundary of the parameter space of the full model (e.g., Self and Liang 1987). This is the case for the test of rate constancy among sites, where the null hypothesis is equivalent to the gamma-distribution model with $\alpha = \infty$. The likelihood differences for the test calculated in this paper were all very large, so the inaccuracy of the χ^2 approximation is not expected to change the conclusions of the tests. Nevertheless, when the likelihood differences are close to the critical values of the χ^2 distribution, the effect needs to be considered.

In this regard, it is worthwhile to note that Tateno et al. (1994) and Gaut and Lewis (1995) considered the test of positivity of interior branch lengths in the maximum likelihood topology as an evaluation of the statistical significance of the estimated tree, as suggested by, e.g., Felsenstein (1988), and found that the χ^2 approximation to the likelihood ratio statistic often produced misleading results. The finding, however, reflects more the inaccurate interpretation of the hypothesis being tested than the poor performance of the χ^2 approximation. Yang (1994b) noted that interior branch lengths in wrong trees as well as those in the true tree can be significantly greater than zero in real data samples, and the positivity of interior branch lengths in the maximum-likelihood tree has nothing to do with the reliability of the topology (see also Goldman and Yang 1994; Yang et al. 1995). The theoretical distribution of the likelihood ratio statistic for testing the positivity of interior branch lengths can be derived by computer simulation when the χ^2 approximation is unreliable (Goldman 1993), but the problem lies in the null hypothesis being misinterpreted (see also Kishino and Hasegawa 1989).

‘‘Saturation’’ of Substitutions

At the third codon positions of the hominoid mtDNA data, about 3.7 substitutions per site have occurred along the tree, and the transition/transversion rate ratio is about 52 (Table 3b). It is often suggested that such data were ‘‘saturated’’ by substitutions (especially transitions) and should not be used in phylogenetic analysis. When sequences are quite different, parsimony reconstructions of ancestral sequences and analyses based on them will be unreliable. In a pairwise comparison, the expected distance between two sequences is a sum of branch lengths along a path in the tree. Pairwise distances can become large even with moderately different sequences and estimates of large distances involve large sampling errors. It is, however, apparent in the analyses of this paper that data at the third codon position contain more information than the first or second codon positions. For example, estimates of parameters for the third position involve smaller sampling errors than those for the first and second positions (e.g., Table 6a), and the third position produced the correct tree topology under the assumption of a molecular clock, while the first and second positions failed to do so. The notion of ‘‘saturation’’ thus appears to depend on the analytical method and should be taken with caution.

Acknowledgments. I thank Naoko Takezaki for drawing Fig. 2 and for many helpful comments on an earlier version of the manuscript. I also thank David Penny and John Huelsenbeck for many useful comments. This study was supported by a grant from the National Science Foundation of China to Z.Y. and by NSF and NIH grants to Masatoshi Nei.

References

Adachi J, Hasegawa M (1995) Improved dating of the human/chimpanzee separation in the mitochondrial DNA tree: heterogeneity among amino acid sites. J Mol Evol 40:622–628

Agresti A, Yang MC (1987) An empirical investigation of some effects of sparseness in contingency tables. Comput Stat Data Anal 5:9–21

Bull JJ, Huelsenbeck JP, Cunningham CW, Swofford DL, Waddell P (1993) Partitioning and combining data in phylogenetic analysis. Syst Biol 42:384–397

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

Felsenstein J (1988) Phylogenies from molecular sequences: inference and reliability. Annu Rev Genet 22:521–565

Gaut BS, Lewis PO (1995) Success of maximum likelihood phylogeny inference in the four-taxon case. Mol Biol Evol 12:152–162

Goldman N (1993) Statistical tests of models of DNA substitution. J Mol Evol 36:182–198

Haberman SJ (1977) Log-linear models and frequency tables with small expected cell counts. Ann Stat 5:1148–1169

Hasegawa M, Kishino H, Yano T (1985) Dating the human-ape splitting by a molecular clock of mitochondrial DNA. J Mol Evol 22:160–174

Horai S, Satta Y, Hayasaka K, Kondo R, Inoue T, Ishida T, Hayashi S, Takahata N (1992) Man’s place in Hominoidea revealed by mitonchondrial DNA genealogy. J Mol Evol 35:32–43

Kishino H, Hasegawa M (1989) Evaluation of maximum likelihood estimate of the evolutionary tree topologies from DNA sequence data, and the branching order in Hominoidea. J Mol Evol 29:170–179

McCullagh P, Nelder JA (1989) Generalized linear models. Chapman and Hall, London

Pilbeam D (1986) Distinguished lecture: hominoid evolution and hominoid origins. Am Anthropol 88:295–312

Reeves JH (1992) Heterogeneity in the substitution process of amino acid sites of proteins coded for by mitochondrial DNA. J Mol Evol 35:17–31

Searle SR (1971) Linear models. Wiley, New York

Self SG, Liang K-Y (1987) Asymptotic properties of maximum likelihood estimators and likelihood ratio tests under nonstandard conditions. J Am Stat Assoc 82:605–610

Swofford DL (1991) When are phylogeny estimations from molecular and morphological data incongruent? In: Miyamoto MM, Cracraft J (eds) Phylogenetic analysis of DNA sequences. Oxford University Press, New York, pp 294–333

Takezaki N, Rzhetsky A, Nei M (1995) Phylogenetic test of the molecular clock and linearized trees. Mol Biol Evol 12:823–833

Tateno Y, Takezaki N, Nei M (1994) Relative efficiencies of the maximum-likelihood, neighbor-joining, and maximum-parsimony methods when substitution rate varies with site. Mol Biol Evol 11:261–277

Wakeley J (1993) Substitution rate variation among sites in hypervariable region 1 of human mitochondrial DNA. J Mol Evol 37:613–623

Wakeley J (1994) Substitution rate variation among sites and the estimation of transition bias. Mol Biol Evol 11:436–442

Yang Z (1993) Maximum likelihood estimation of phylogeny from DNA sequences when substitution rates differ over sites. Mol Biol Evol 10:1396–1401

Yang Z (1994a) Estimating the pattern of nucleotide substitution. J Mol Evol 39:105–111

Yang Z (1994b) Statistical properties of the maximum likelihood method of phylogenetic estimation and comparison with distance matrix methods. Syst Biol 43:329–342

Yang Z (1994c) Maximum likelihood phylogenetic estimation from DNA sequences with variable rates over sites: approximate methods. J Mol Evol 39:306–314

Yang Z, Goldman N, Friday AE (1994) Comparison of models for nucleotide substitution used in maximum likelihood phylogenetic estimation. Mol Biol Evol 11:316–324

Yang Z, Goldman N, Friday AE (1995) Maximum likelihood trees from DNA sequences: a peculiar statistical estimation problem. Syst Biol 44:384–399