On the Use of Nucleic Acid Sequences to Infer Early Branchings in the Tree of Life

*Ziheng Yang**,† *and Dave Roberts**

*Department of Zoology, The Natural History Museum, United Kingdom, and †College of Animal Science, Beijing Agricultural University, China

Simplifying assumptions made in various tree reconstruction methods—notably rate constancy among nucleotide sites, homogeneity, and stationarity of the substitutional processes—are clearly violated when nucleotide sequences are used to infer distant relationships. Use of tree reconstruction methods based on such oversimplified assumptions can lead to misleading results, as pointed out by previous authors. In this paper, we made use of a (discretized) gamma distribution to account for variable rates of substitution among sites and built models that allowed for unequal base frequencies in different sequences. The models were nonhomogeneous Markov-process models, assuming different patterns of substitution in different parts of the tree. Data of the small-subunit rRNAs from four species were analyzed, where base frequencies were quite different among sequences and rates of substitution were highly variable at sites. Parameters in the models were estimated by maximum likelihood, and models were compared by the likelihood-ratio test. The nonhomogeneous models provided significantly better fit to the data than homogeneous models despite their involvement of many parameters. They also appeared to produce reasonable estimation of the phylogenetic tree; in particular, they seemed able to identify the root of the tree.

Introduction

Nucleotide sequences have been widely used in reconstructing evolutionary trees that represent relationships among living species. The early evolution of life was unicellular and left very little fossil trace and no residual morphological characters. Molecular sequences are the only source of data from which their distant relationships can be inferred. Small-subunit rRNA (ss rRNA) sequences are the most widely used, for example, to infer the origin of metazoa (Field et al. 1988; Lake 1990; Wainright et al. 1993); the early evolution of eukaryotes (Sogin et al. 1989; Sogin 1991); and the earliest splittings among archaebacteria, eubacteria, and eukaryotes since the origin of life (Woese 1987; Zillig et al. 1989; Woese et al. 1990; Sogin 1991; Rivera and Lake 1992; Cavalier-Smith 1993; Forterre et al. 1993; Olsen et al. 1994).

The use of nucleotide sequences such as ss rRNAs to infer deep branchings in the tree of life involves assumptions made in various tree reconstruction methods, which, while perhaps tenable for closely related sequences, are unacceptable when distantly related species are compared. To analyze distant relationships, genes that perform fundamental roles in life and exist in all organisms must be used. Rates of nucleotide substitution at different sites are highly variable in such genes because of the existence of structural and functional domains or "variable" and "conservative" regions in the gene. Most tree reconstruction methods, either explicitly or implicitly, assume a constant rate for all sites. Ignoring rate variation among sites has been found to affect drastically certain aspects of phylogenetic analysis, leading, for example, to severe underestimation of branch lengths and of the transition:transversion rate ratio (see, e.g., Gillespie 1986; Takahata 1991; Wakeley 1994; Yang et al. 1994). Presumably, estimation of the tree topology will also be affected. Many attempts have been taken to alleviate this problem; for example, formulas have been suggested for estimating the distance between two sequences with substitution rates for sites assumed to follow a gamma distribution (see, e.g., Jin and Nei 1990; Li et al. 1990; Tamura and Nei 1993). The gamma-distribution model was found to fit real data quite well (see, e.g., Wakeley 1993; Yang et al. 1994) and has been extended to a maximum-likelihood (ML) joint analysis of all sequences (Yang 1993). This method, however, involved very intensive computation, and a "discrete gamma" model has been suggested by Yang (1994) whereby several equal-probability categories were used

Key words: models, maximum likelihood, unequal base frequencies, G+C content, DNA sequences, molecular systematics.

Address for correspondence: Ziheng Yang, Institute of Molecular Evolutionary Genetics, The Pennsylvania State University, 328 Mueller Laboratory, University Park, Pennsylvania 16802. E-mail: yang@imeg.bio.psu.edu

Mol. Biol. Evol. 12(3):451–458. 1995.

0737-4038/95/1203-0010$02.00

to approximate the (continuous) gamma distribution, with the mean of each category used to represent all rates in the category. Analyses of several data sets suggested that four categories were usually sufficient to provide an optimum or near-optimum fit by the model to data and also a satisfactory approximation to the continuous distribution (Yang 1994).

Another common assumption is that the processes of nucleotide substitution are time-homogeneous and stationary; that is, substitutions follow the same (stochastic) pattern in different lineages, and the overall base frequencies do not change over time. This is unrealistic when sequences from distantly related species are compared. An obvious, but nevertheless very sensitive, indication of a violation of this assumption is the observation that base frequencies in different species are quite different. It has been suggested that tree topologies obtained from analysis of ss rRNAs can be misleading when base frequencies are very different among species (Loomis and Smith 1990; Hasegawa and Hashimoto 1993; Hasegawa et al. 1993). Where genomes have independently acquired similar base compositions, parsimony methods tend to suggest a wrong tree by grouping together sequences with similar base frequencies rather than with similar genetic background (Lockhart et al. 1992; Steel et al. 1993). Embley et al. (1993) observed that the problem of unequal base frequencies is usually confounded with the problem of long branches in the tree, which is known to mislead parsimony methods (see, e.g., Felsenstein 1978), and that by sampling taxa to break long branches, the problem of unequal base frequencies can be alleviated to some extent.

Perceiving the pitfalls of blind application of tree reconstruction methods to ss rRNAs, Hasegawa et al. (1993) argue that trees constructed from protein sequences may be more reliable. Sometimes, even when the coding DNA sequences are available, they are translated into proteins for analysis, because the frequencies of the amino acids are much more homogeneous among species than those of nucleotides or codons (Loomis and Smith 1990; Adachi et al. 1993). This practice is able to remove much of the "noise" in the data, but the loss of information due to collapsing character states is extraordinary.

A third problem is that of alignment, which becomes very serious when the sequences are distantly related. The variable regions of ss rRNAs are difficult to align; in these regions, apart from the many substitutions that have occurred, insertions and deletions are also commonplace. Unequal base frequencies in different sequences add further difficulty to alignment. Ideally, a model that allows for insertions and deletions as well as substitutions will make use of the extra information provided by the existence of gaps in the sequences and will presumably lead to more reliable estimation of phylogeny (Thorne et al. 1991, 1992). At present, a model like this appears very complicated, especially when the problems of variable rates among sites and unequal base frequencies among species are to be addressed. For this reason, the problem of alignment has been excluded from this study, and the data were assumed to be correctly aligned sequences with gaps removed before analysis.

Phylogenetic trees derived from the ss rRNAs may be misleading not because of any incorrect information contained in the data but because of inadequate analytical methods based on unrealistic assumptions. Methods for testing models of nucleotide substitution have been described by Ritland and Clegg (1987), Reeves (1992), and Goldman (1993*a*, 1993*b*); and their accuracy in the context of phylogenetic estimation has been examined by Goldman (1993*a*, 1993*b*). As the assumptions made are found to be of importance to phylogenetic analysis (Yang et al. 1994), it is not surprising that analyses based on totally wrong assumptions may lead to spurious results.

In this study, the discrete gamma model (Yang 1994) was used to allow variable rates for sites, and the problem of unequal base frequencies among sequences was addressed. The maximum-likelihood (ML) framework of phylogenetic estimation (Felsenstein 1981) was adopted, and nonhomogeneous Markov-process models which allow different patterns of substitution along different branches of the tree, and thus unequal base frequencies in different sequences, were constructed. In a sense, the processes that have generated the noise in the data (unequal base frequencies) are being modeled in the hope that the phylogenetic information can be extracted, while the practice of translating DNA sequences into proteins, where feasible, is equivalent to ignoring part of the data in order to achieve a higher information: noise ratio. An obvious problem with the nonhomogeneous models is that they involve many parameters so that estimation may be inefficient. Preliminary examination of this problem will be performed by analyzing a real data set.

Data and Methods

Data

The aligned ss rRNA sequences of *Sulfolobus solfatarius* (an archaebacterium), *Halobacterium salinarium* (another archaebacterium, a synonym of *H. halobium* by which it may be found in the international databases), *Escherichia coli* (a eubacterium), and *Homo sapiens* (a eukaryote) were obtained from W. Navidi. The data contain 1,352 nucleotides in each sequence and have been analyzed by several authors for different

purposes (Navidi et al. 1991; Yang 1994; Yang et al. 1994). The base frequencies are listed in table 1, and they are seen to differ among sequences.

Pattern of Nucleotide Substitution

A locally homogeneous Markov process was used to model nucleotide substitution along a branch in the tree, but different processes were allowed for different branches. The basic model was that of Hasegawa et al. (1985), by which the rate of nucleotide i changing into nucleotide $j (j \neq i)$ is

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

where π_j is the equilibrium frequency of nucleotide j, with $\sum \pi_j = \pi_T + \pi_C + \pi_A + \pi_G = 1$. The diagonals of the rate matrix $\mathbf{Q} = \{Q_{ij}\}$ are determined by the mathematical restriction that row sums of $\mathbf{Q}$ are zero; $-Q_{ii} = \sum_{j \neq i} Q_{ij}$ is then the rate of substitution of nucleotide i. The matrix $\mathbf{Q}$ is multiplied by a scale factor such that the expected rate of substitution at equilibrium is $-\sum \pi_i Q_{ii} = 1$. This means that time t, or the branch length in a tree, is measured by the expected number of substitutions per site accumulated during the time period or along the branch. The transition probability matrix for the branch (with length t) is then $\mathbf{P}(t) = \exp(\mathbf{Q}t)$. The model will be referred to as HKY, and, when combined with the gamma or discrete gamma distribution for variable rates among sites, as HKY+Γ or HKY+dG.

Nonhomogeneous Process Models

Different frequency parameters (π_i's in eq. [1]) are allowed for different branches in the tree so that base frequencies can drift toward different values in different lineages. The models may be considered special cases of Barry and Hartigan's (1987) parameter-rich model, whereby one whole-rate matrix (including 11 free parameters) was assigned for each branch. To get some feel for the cost of using many parameters on the one hand and the fit of the model to data on the other, two models were constructed concerning the frequency parameters, referred to as N1 and N2. Model N2 is the more general, in which one set of frequency parameters (three free parameters) is assigned for each branch. Including the initial base frequencies at the root of the tree, this model involves as many sets of frequency parameters as the number of nodes in the (rooted) tree; for a bifurcating tree with s species, this number is $2s - 1$. As the interior branches are usually short, one set of frequency parameters was used for all interior branches in model N1. Including one set for the root and one set for each of the branches leading to the end nodes, model N1 involves $s + 2$ sets of frequency parameters and is a special case of N2.

The likelihood function for given values of parameters and tree topology can be calculated following Felsenstein (1981) (see also Barry and Hartigan 1987) for models assuming a single rate for all sites, and following Yang (1994) for models that assume the (discrete) gamma distribution for rates over sites.

Removal of the homogeneity and stationarity assumption created two new problems. First, the placement of the root in a tree changes the likelihood, and therefore rooted trees should be considered, as implied above. This problem was ignored by Barry and Hartigan (1987). Second, as the process of substitution is not at equilibrium and base frequencies change with time, the branch length (t) as defined above is only an approximation to the real expected number of substitutions accumulated along the branch, which is an average over variable base frequencies. In this study, we use t as the (approximate) branch length, as this approximation will slightly affect estimates of branch lengths only, and estimates of other parameters or calculation of the likelihood are unaffected.

Results

Homogeneous Models

Table 2 lists results obtained from using models with the homogeneity and stationarity assumptions. The HKY model of substitution is used, either assuming a single rate for all sites (HKY) or gamma-distributed rates

Table 1
Nucleotide Frequencies in ss rRNAs (1,352 nucleotides) of Four Species Analyzed in this Paper

Species	T	C	A	G	G + C
1. *Sulfolobus solfatarius*	0.1428	0.2825	0.2249	0.3499	0.6324
2. *Halobacterium salinarium*	0.1790	0.2567	0.2345	0.3299	0.5866
3. *Escherichia coli*	0.1990	0.2367	0.2507	0.3136	0.5503
4. *Homo sapiens*	0.2182	0.2411	0.2485	0.2922	0.5333
Average	0.1847	0.2543	0.2396	0.3214	0.5757

Table 2
Log-Likelihood Values and Estimates of Parameters Obtained from Models Assuming Homogeneity and Stationarity of the Substitution Processes

	HKY		HKY + Γ			HKY + dG		
TREE	$\ell - \ell_{max}$	$\hat{\kappa}$	$\ell - \ell_{max}$	$\hat{\kappa}$	$\hat{\alpha}$	$\ell - \ell_{max}$	$\hat{\kappa}$	$\hat{\alpha}$
T_0: (1234)	−248.40	1.83	−195.22	2.46	0.77	−195.42	2.46	0.77
T_1: ((12)34)	**−240.67**	**1.84**	−195.20	2.45	0.78	−195.39	2.44	0.78
T_2: ((13)24)	−244.86	1.81	−194.43	2.41	0.80	−194.51	2.41	0.79
T_3: ((14)23)	−244.91	1.83	**−193.83**	**2.43**	**0.80**	**−193.85**	**2.42**	**0.79**

NOTE.—The ss rRNAs of *Sulfolobus solfatarius* (1), *Halobacterium salinarium* (2), *Escherichia coli* (3), and *Homo sapiens* (4) are analyzed. $\ell_{max} = -5{,}591.06$ is the upper limit of the log likelihood (Navidi et al. 1991; Goldman 1993*a*). Parameter κ is the transition:transversion rate ratio (see eq. [1]), and α is the shape parameter of the gamma (Γ) or discrete gamma (dG) distribution for rates among sites. Parameters in the models are estimated by maximum likelihood for each of the four (unrooted) tree topologies, and estimates of branch lengths are not shown. T_0 is the star tree. Estimates of the frequency parameters obtained from the four trees are identical at the third decimal point; these are $\hat{\pi}_T = 0.205$, $\hat{\pi}_C = 0.265$, $\hat{\pi}_A = 0.224$, and $\hat{\pi}_G = 0.306$ for the HKY model; $\hat{\pi}_T = 0.203$, $\hat{\pi}_C = 0.270$, $\hat{\pi}_A = 0.217$, and $\hat{\pi}_G = 0.310$ for the HKY + Γ and HKY + dG models. Results for the ML trees under the models are shown in boldface type.

among sites (HKY+Γ). In the discrete gamma model (HKY+dG), four categories of rates are used to approximate the gamma distribution (Yang 1994). The shape parameter α of the gamma distribution is inversely related to the extent of rate variation over sites; $\alpha = \infty$ corresponds to the case of a single rate for all sites. HKY is thus a special case of HKY+Γ or HKY+dG.

In all three models, the substitution processes were assumed to be homogeneous and stationary, and one κ and one set of frequency parameters were assumed for all sequences (branches). The existence of a molecular clock (i.e., rate constancy among lineages) was not assumed, and as HKY is a reversible process model, the root of the tree cannot be identified (Felsenstein 1981). All parameters in the models were estimated from the data for each (unrooted) tree topology. The ML estimates of π_i's are quite different from the averages of the observed frequencies; for example, using the average observed frequencies (table 1) in the HKY+Γ model gives log-likelihood values −202.62, −202.57, −201.69, and −201.15 for tree topologies T_0, T_1, T_2, and T_3, respectively (compare table 2). Rate variation over sites can be seen from the tremendous improvement in likelihood upon adding the α parameter of the gamma distribution (comparison between HKY and HKY+Γ or HKY+dG). For example, using tree topology T_3, we can compare HKY and HKY+Γ by the likelihood-ratio test to test for rate constancy among sites, which means comparison of $2\Delta\ell = 2(-193.83 - [-244.91]) = 102.16$ with $\chi^2_{1,1\%} = 6.63$, and the difference is obviously significant. The same conclusion is drawn if T_1 or T_2 is used in the comparison instead of T_3, or if HKY+dG is used instead of HKY+Γ (table 2).

T_1 was the best tree by the HKY model, while assuming the gamma distribution of rates over sites favored T_3 (HYK+Γ and HKY+dG). The likelihood values for different trees were worryingly similar under the HKY+Γ and HKY+dG models. It was also noted that the performance of the HKY+dG model was quite good relative to HKY+Γ, with respect both to the fit to data reflected in the likelihood values and to the approximation to the continuous distribution reflected in the estimates of the α parameter. The discrete gamma model was used in later analysis in place of the continuous gamma.

Nonhomogeneous Models

In a preliminary analysis of the ss rRNA data, a model that assumed one κ for each branch in the tree was compared with another that assumed one κ for all branches in the tree. The likelihood values obtained for these two models were not significantly different, indicating that the transition:transversion rate ratio is more or less the same in different parts of the tree although the sequences are drifting toward different base frequencies; at any rate, the ratio is not much larger than one (compare Yang et al. 1994). Later analyses were performed assuming one κ for the whole tree.

Results obtained from models that do not assume homogeneity and stationarity of substitution (models N1 and N2) are listed in table 3. The 15 (rooted) bifurcating trees are classified into three groups according to their unrooted topology; for example, trees T_{11}, T_{12}, T_{13}, T_{14}, and T_{15} in table 3 have the same unrooted topology (i.e., T_1 in table 2). Compared to HKY (table 2), the HKY+N2 model (table 3) involves one extra branch (length) due to the addition of the root and $3 \times (2s - 1 - 1) = 18$ extra frequency parameters. The homogeneity and stationarity assumptions are clearly rejected; for example, using the likelihood values of T_{11} (table 3) and T_1 (table 2), we compare $2\Delta\ell = 2 \times (-175.13 - [-240.67]) = 131.08$ with $\chi^2_{19,1\%} = 36.19$,

Table 3
Log-Likelihood Values and Parameter Estimates Under Nonhomogenous Models of Nucleotide Substitution

	HKY + N2		HKY + dG + N2			HKY + dG + N1		
TREE	$\ell - \ell_{max}$	$\hat{\kappa}$	$\ell - \ell_{max}$	$\hat{\kappa}$	$\hat{\alpha}$	$\ell - \ell_{max}$	$\hat{\kappa}$	$\hat{\alpha}$
T_0: (1234)	−205.46	1.88	−149.70	2.83	0.65	−149.70	2.83	0.65
T_{11}: (((12)4)3)	**−175.13**	**1.97**	*−126.30*	*2.74*	*0.76*	*−127.48*	*2.77*	*0.73*
T_{12}: (((12)3)4)	−193.34	1.94	−136.20	2.87	0.66	−138.33	2.81	0.66
T_{13}: (((34)2)1)	−191.06	1.90	−146.04	2.60	0.74	−146.47	2.74	0.69
T_{14}: (((34)1)2)	−188.85	1.93	−141.14	2.63	0.72	−141.98	2.75	0.69
T_{15}: ((12)(34))	−192.64	1.92	−143.93	2.90	0.65	−143.93	2.89	0.65
T_{21}: (((24)1)3)	*−180.27*	*1.95*	*−126.58*	*2.71*	*0.75*	*−126.74*	*2.72*	*0.75*
T_{22}: (((13)2)4)	−198.60	1.90	−136.63	2.84	0.65	−137.72	2.85	0.64
T_{23}: (((13)4)2)	−192.10	1.90	−139.83	2.66	0.73	−141.98	2.75	0.69
T_{24}: (((24)3)1)	−196.37	1.87	−145.31	2.67	0.72	−145.88	2.69	0.71
T_{25}: ((13)(24))	−199.68	1.86	−147.04	2.78	0.66	−147.17	2.77	0.67
T_{31}: (((14)2)3)	*−177.87*	*1.96*	**−123.43**	**2.72**	**0.75**	**−123.67**	**2.72**	**0.75**
T_{32}: (((14)3)2)	−191.96	1.91	−138.29	2.66	0.73	−138.66	2.67	0.73
T_{33}: (((23)4)1)	−195.91	1.88	−145.29	2.72	0.70	−146.40	2.74	0.69
T_{34}: (((23)1)4)	−198.42	1.91	−136.61	2.80	0.65	−138.55	2.87	0.64
T_{35}: ((14)(23))	−200.60	1.87	−144.14	2.72	0.69	−148.33	2.69	0.68

NOTE.—The ss rRNAs of *Sulfolobus solfatarius* (1), *Halobacterium salinarium* (2), *Escherichia coli* (3), and *Homo sapiens* (4) are analyzed. T_0 is the star tree, and other multifurcating trees are not evaluated. Results for the ML trees are given in boldface type, while those for the best tree in each group corresponding to the same unrooted topology are listed in italicized type. A four-category discrete gamma (dG) model is used to account for variable rates over sites in the dG models. One set of frequency parameters is assumed for each branch in the tree by the N2 model, while the N1 model differs from N2 by assuming one set of frequency parameters for all the interior branches. Estimates of the frequency parameters and of branch lengths are not presented.

and the difference is significant. Similar results are obtained if other reasonable tree topologies such as T_{21} or T_{31} are used, or if rate variation over sites has been taken into account (comparison between HKY+dG+N2 with HKY+dG). In sum, both rate variation among sites and nonhomogeneity of substitution are characteristics of the evolutionary processes of these sequences. The HKY+dG+N1 model has three fewer frequency parameters than HKY+dG+N2 but very similar log-likelihood values. Results obtained from HKY+N1 (not shown) are also very similar to those obtained from HKY+N2; for example, the three best trees under HKY+N1 are also T_{11}, T_{31}, and T_{21}, with log likelihoods of −177.76, −178.01, and −180.76, respectively. The reason appears to be that one of the two interior branches in any tree topology is very short, and not much information exists in the data concerning the pattern of substitution along the short branch.

Estimates of parameters that are common to all tree topologies; that is, κ and α are remarkably similar for different tree topologies. Furthermore, the transition: transversion rate ratio is underestimated under the HKY model which ignores rate variation over sites. These were observed and discussed for other models or data sets (Wakeley 1994; Yang et al. 1994). Branch lengths (not shown) are also underestimated when rate variation over sites is ignored.

It is noteworthy that different rooted trees sharing the same unrooted topology have quite different likelihood values (table 3). The three tree topologies that suggest first separation of *Escherichia coli* from other species—T_{11}, T_{21}, and T_{31},—have likelihood values much higher than others. This seems to suggest that the data contain considerable information concerning the position of the root (the earlist splitting) even though the topology may be uncertain. Notably the likelihood difference between the ML tree and the second best tree under the HKY+dG+N1 or HKY+dG+N2 models (table 3) is larger than that between the ML tree and the star tree under the HKY+dG model (table 2). This seems to suggest that the application of nonhomogeneous models to sequences with different base frequencies not only allows the root of the tree to be located but also leads to better power in discriminating among tree topologies.

The best tree under HKY+dG+N2 (fig. 1) separates the species in the order *E. coli, Halobacterium salinarium, Sulfolobus solfatarius,* and *Homo sapiens,* suggesting that *E. coli,* which represents eubacteria, separates first from the other species and that the two archaebacterial species (*H. salinarium* and *S. solfatarius*) are not monophyletic. The position of the root of the universal tree of life is widely, if not universally, accepted as being within the branch of eubacteria. Support for this position

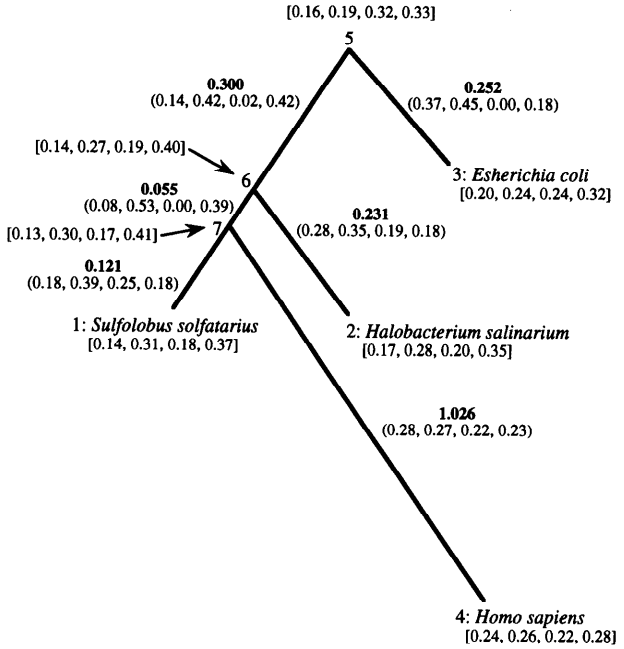

FIG. 1.—The maximum-likelihood tree and estimates of parameters for ss rRNAs under the HKY+dG+N2 model. One κ parameter in the HKY model is assumed for the whole tree, with estimate $\hat{\kappa} = 2.72 \pm 0.30$. One set of frequency parameters is assumed for each branch (N2). A discrete gamma model (dG) is used to describe variable rates among sites, with $\hat{\alpha} = 0.75 \pm 0.10$. The log likelihood for this tree is $\ell - \ell_{max} = -5{,}714.49 - (-5{,}591.06) = -123.43$. Branch lengths, approximately measured by the expected numbers of nucleotide substitutions per site, are shown in boldface type. Numbers in parentheses are estimates of the frequency parameters in the HKY model (eq. [1]) for the branch, while those in brackets are the base frequencies in sequences at the nodes of the tree, estimated from the model. For example, the base frequencies at the root of the tree (node 5) are estimated as 0.16 (*T*), 0.19 (*C*), 0.32 (*A*), and 0.33 (*G*). These estimates of frequency parameters, however, involve large sampling errors and do not appear reliable.

has been derived by using gene duplications which are believed to have occurred before the separation of these lineages (Gotgarten et al. 1989; Iwabe et al. 1989). In contrast, the nonhomogeneous models identified the root of the tree by using one single gene. Although the earliest separation of *E. coli* seems to be strongly supported as tree topologies T_{11}, T_{21}, and T_{31} have much higher likelihood values than other tree topologies, the relationship among the remaining three species is much less certain, in that the likelihood values for the three trees T_{11}, T_{21}, and T_{31} are similar (table 3). No attempt is made here to evaluate the reliability (sampling error) of the ML tree (T_{31}), and we suggest that the present results do not contribute to the debate over monophyly of the archaebacteria (Hoffman 1992). The eukaryotic lineage, represented by *Homo sapiens,* is placed on the longest branch indicating greatest divergence; this is true as long as one of the three best trees (i.e., T_{11}, T_{21}, and T_{31}) is the true tree. This might reflect two important events in eukaryotic evolution: the appearance of chromosomes and mitotic division, and the evolution of sexual reproduction. These two events might affect the rate of mutation within the lineage, particularly within the early eukaryotes.

Estimates of branch lengths and frequency parameters for the branches are shown in figure 1 for the ML tree under the HKY+dG+N2 model (i.e., T_{31}). The expected base frequency distribution at a node of the tree is calculated from the formula $\mathbf{p}_t = \mathbf{p}_0 \cdot \mathbf{P}(t)$, where the row vectors $\mathbf{p}_0$ and $\mathbf{p}_t$ are the base frequency distributions at the start and end of the branch, respectively. We note that parameters in a model are estimated with different levels of accuracy. Poor estimates usually have large sampling variances and are sensitive to small perturbations in the model or data. Because of the parameter richness of the nonhomogeneous models, it is important to find out which parameters are reliably estimated and which are not. We have examined this problem by calculating the standard errors of the parameter estimates, by comparing estimates under the HKY+dG+N1 and HKY+dG+N2 models, and by comparing estimates obtained from different tree topologies. Either of the two models may be considered a slight variation of the other, and so are some of the tree topologies such as T_{11}, T_{21}, and T_{31}. We note that estimates of branch lengths are quite stable no matter which of the two models is assumed; their sampling variances are also comparable to those of branch length estimates under the homogeneous models. One exception is for the two branches around the root of the tree; using figure 1 as an example, estimates of lengths for branches 5-3 and 5-6 can be quite different by the two models although their sum is almost the same; although the root of the tree is quite certain, the exact position of the root is not reliably estimated. Likelihood values and estimates of κ, α, and other branch lengths appear to be quite reliable by this evidence. The frequency parameters for the branches involve large sampling variances and may have quite different estimates under the two models considered. Their estimates are the least reliable.

Discussion

Use of the nonhomogeneous models in this study suggests that the ss rRNAs can lead to reasonable estimation of phylogeny despite the fact that base frequencies are quite different in different species. The results in table 3 also suggest that it may be possible to identify the root of the tree even though the topology is uncertain. If this can be confirmed with more data sets, methods that infer rooted trees should be considered more seriously.

As mentioned before, the N2 model involves $3 \times (2s - 1)$ frequency parameters, besides branch lengths in the tree and parameters such as κ and α which are common to all tree topologies. With s larger than 4 or 5, this means many parameters. The N1 model, by using one set of frequency parameters for all interior branches, reduces the number of frequency parameters to $3 \times (s + 2)$. However, this restriction may be too unrealistic when there are many (long) interior branches. We note that even with four species, the nonhomogeneous models still pose computational problems, especially if the sequences are short; it is difficult to locate the optimum values of the frequency parameters during the iteration. The models do not appear to be usable for data of more than five sequences, and more practical methods are needed. When some species in the data have very similar base frequencies and are known to belong to a monophyletic group (which suggests that within the group the patterns of substitution are more or less the same), we may use one set of frequency parameters for all branches within the group. Another possibility may be to take the frequencies in the rate matrices for different branches as random variables generated from a probabilistic distribution, analogous to the case of using a gamma distribution (with a single parameter α) to describe variable rates among sites rather than estimating one rate parameter for each site. At present, it is unclear how such a distribution for a superprocess of base frequency drift can be constructed. Even within the framework of the nonhomogeneous models of this paper, it may be feasible to obtain approximations to the transition probabilities for the branches in the tree without iteration, using, for example, the parsimony inference of the ancestral sequences; this approach will remove the computational problems of the nonhomogeneous models mentioned above.

It should be noted that the models considered in this paper are all formulated at the level of nucleotide substitution, which is the product of a complicated process driven by many factors, notably mutation, random drift and natural selection. Unequal base compositions in different species may have a strong selectional basis; for example, thermophiles probably have gained a selective advantage in maintaining high G+C content. These factors are, however, very difficult to model, and in our formulation the effects of selection are reflected in the affected substitution rates. Therefore, the locally homogeneous Markov process assumed for one branch in the tree should be interpreted as an average over time of a substitution pattern that varies along the branch. The argument of Barry and Hartigan (1987) suggests that it may be impossible to distinguish using sequence data the average of a variable substitution pattern along a branch from a constant pattern. For data to which the nonhomogeneous models are expected to apply, the interpretation of a variable pattern along a branch is biologically more reasonable. Instead of using a rate matrix $\mathbf{Q}$ for a branch as described in equation (1), the models may as well be formulated using only the matrix $\mathbf{P}(t)$ of transition probabilities along the branch; the rate matrix $\mathbf{Q}$ may be considered a way of placing restrictions on the structure of $\mathbf{P}(t)$ to reduce the number of parameters.

Acknowledgment

We wish to thank Dr. N. Takahata and two anonymous referees for many useful comments. Dr. M. Nei provided facilities during the revision of the manuscript. This study was partially supported by a grant from the National Science Foundation of China to Z.Y.

LITERATURE CITED

ADACHI, J., Y. CAO, and M. HASEGAWA. 1993. Tempo and mode of mitochondrial DNA evolution in vertebrates at the amino acid sequence level: rapid evolution in warm-blooded vertebrates. J. Mol. Evol. **36**:270–281.

BARRY, D., and J. A. HARTIGAN. 1987. Statistical analysis of hominoid molecular evolution. Statist. Sci. **2**:191–210.

CAVALIER-SMITH, T. 1993. Kingdom Protozoa and its 18 phyla. Microbiol. Rev. **57**:953–994.

EMBLEY, T. M., R. H. THOMAS, and R. A. D. WILLIAMS. 1993. Reduced thermophilic bias in the 16S rDNA sequence from *Thermus ruber* provides further support for a relationship between *Thermus* and *Deinococcus*. Syst. Appl. Microbiol. **16**:25–29.

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

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

FIELD, K. G., G. J. OLSEN, D. J. LANE, S. J. GIOVANNONI, M. T. GHISELIN, E. C. RAFF, N. R. PACE, and R. A. RAFF. 1988. Molecular phylogeny of the animal kingdom. Science **239**:748–753.

FORTERRE, P., N. BENACHENHOU-LAHFA, F. CONFALONIERI, M. DUGUET, C. ELIE, and B. LABEDAN. 1993. The nature of the last universal ancestor and the root of the tree of life, still open questions. BioSystems **28**:15–32.

GILLESPIE, J. H. 1986. Rates of molecular evolution. Ann. Rev. Ecol. Syst. **17**:637–65.

GOLDMAN, N. 1993*a*. Statistical tests of models of DNA substitution. J. Mol. Evol. **36**:182–198.

———. 1993*b*. Simple diagnostic statistical tests of models for DNA substitution. J. Mol. Evol. **37**:650–661.

GOTGARTEN, J. P., H. KIBAK, P. DITTRICH, L. TAIZ, E. J. BOWMAN, B. J. BOWMAN, M. F. MANOLSON, R. J. POOLE, T. DATE, T. OSHIMA, J. KONISHI, K. DENDA, and M. YOSHIDA. 1989. Evolution of the vacuolar H^+-ATPase: implications for the origin of eukaryotes. Proc. Natl. Acad. Sci. USA **86**:6661–6665.

HASEGAWA, M., and T. HASHIMOTO. 1993. Ribosomal RNA trees misleading? Nature **361**:23.

HASEGAWA, M., T. HASHIMOTO, E. OTAKA, J. ADACHI, N. IWABE, and T. MIYATA. 1993. Early divergences in the evolution of eukaryotes: ancient divergence of *Entamoeba* that lacks mitochondria revealed by protein sequence data. J. Mol. Evol. **36**:380–388.

HASEGAWA, M., H. KISHINO, and T. YANO. 1985. Dating of the human-ape splitting by a molecular clock of mitochondrial DNA. J. Mol. Evol. **22**:160–174.

HOFFMAN, M. 1992. Researchers find organisms they can really relate to. Science **257**:32.

IWABE, N., K. KUMA, M. HASEGAWA, S. OSAWA, and T. MIYATA. 1989. Evolutionary relationships of archaebacteria, eubacteria and eukaryotes infered from phylogenetic trees of duplicated genes. Proc. Natl. Acad. Sci. USA **86**:9355–9359.

JIN, L., and M. NEI. 1990. Limitations of the evolutionary parsimony method of phylogenetic analysis. Mol. Biol. Evol. **7**:82–102.

LAKE, J. A. 1990. Origin of the Metazoa. Proc. Natl. Acad. Sci. USA **87**:763–766.

LI, W.-H., M. GOUY, P. M. SHARP, C. O'HUIGIN, and Y.-W. YANG. 1990. Molecular phylogeny of rodentia, lagomorpha, primates, artidactyla and carnivora and molecular clocks. Proc. Natl. Acad. Sci. USA **87**:6703–6707.

LOCKHART, P. S., C. J. HOWE, D. A. BRYANT, T. J. BEANLAND, and A. W. D. LARKUM. 1992. Substitutional bias confounds inference of cyanelle origins from sequence data. J. Mol. Evol. **34**:153–162.

LOOMIS, W. F., and D. W. SMITH. 1990. Molecular phylogeny of *Dictyostelium discoideum* by protein sequence comparison. Proc. Natl. Acad. Sci. USA **87**:9093–9097.

NAVIDI, W. C., G. A. CHURCHILL, and A. VON HAESELER. 1991. Methods for inferring phylogenies from nucleotide acid sequence data by using maximum likelihood and linear invariants. Mol. Biol. Evol. **8**:128–143.

OLSEN, G. J., C. R. WOESE, and R. OVERBEEK. 1994. The winds of (evolutionary) change: breathing new life into microbiology. J. Bacteriol. **176**:1–6.

REEVES, J. H. 1992. Heterogeneity in the substitution process of amino acid sites of proteins coded for by mitochondrial DNA. J. Mol. Evol. **35**:17–31.

RITLAND, K., and M. T. CLEGG. 1987. Evolutionary analysis of plant DNA sequences. Am. Nat. **130**:S74–S100.

RIVERA, M. C., and J. A. LAKE. 1992. Evidence that eukaryotes and eocyte prokaryotes are immediate relatives. Science **257**: 72–76.

SOGIN, M. L. 1991. Early evolution and the origin of the eukaryotes. Curr. Opin. Genet. Dev. **1**:457–463.

SOGIN, M. L., J. H. GUNDERSON, H. J. ELWOOD, R. A. ALONSO, and D. A. PEATTIE. 1989. Phylogenetic meaning of the kingdom concept: an unusual ribosomal RNA from *Giardia lamblia*. Science **243**:75–77.

STEEL, M. A., P. J. LOCKHART, and D. PENNY. 1993. Confidence in evolutionary trees from biological sequence data. Nature **364**:440–442.

TAKAHATA, N. 1991. Overdispersed molecular clock at the major histocompatibility complex loci. Proc. R. Soc. Lond. B **243**:13–18.

TAMURA, K., and M. NEI. 1993. Estimation of the number of nucleotide substitutions in the control region of mitochondrial DNA in humans and chimpanzees. Mol. Biol. Evol. **10**:512–526.

THORNE, J. L., H. KISHINO, and J. FELSENSTEIN. 1991. An evolutionary model for maximum likelihood alignment of DNA sequences. J. Mol. Evol. **33**:114–124. (Erratum: J. Mol. Evol. **34**:91 [1992].)

———. 1992. Inching toward reliability: an improved likelihood model of sequence evolution. J. Mol. Evol. **34**:3–16.

WAINRIGHT, P. O., G. HINKLE, M. L. SOGIN, and S. K. STICKEL. 1993. Monophyletic origins of the metazoa: an evolutionary link with the fungi. Science **260**:340–342.

WAKELEY, J. 1993. Substitution rate variation among sites in hypervariable region 1 of human mitochondrial DNA. J. Mol. Evol. **37**:613–623.

———. 1994. Substitution rate variation among sites and the estimation of transition bias. Mol. Biol. Evol. **11**:436–442.

WOESE, C. R. 1987. Bacterial evolution. Microbiol. Rev. **51**: 221–271.

WOESE, C. R., L. KANDLER, and M. L. WHEELIS. 1990. Towards a natural system of organisms: proposal for the domains archaea, bacteria, and eucarya. Proc. Natl. Acad. Sci. USA **87**:4576–4579.

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

———. 1994. Maximum likelihood phylogenetic estimation from DNA sequences with variable rates over sites: approximate methods. J. Mol. Evol. **39**:306–314.

YANG, Z., N. GOLDMAN, and A. E. FRIDAY. 1994. Comparison of models for nucleotide substitution used in maximum likelihood phylogenetic estimation. Mol. Biol. Evol. **11**:316–324.

ZILLIG, W., H. P. KLENK, P. PALM, H. LEFFERS, G. PÜHLER, F. GROPP, and R. GARRETT. 1989. Did eukaryotes originate by a fusion event? Endocytobiosis Cell Res. **6**:1–25.

NAOYUKI TAKAHATA, reviewing editor

Received August 12, 1994

Accepted December 13, 1994