Inferring pattern and process: maximum-likelihood implementation of a nonhomogeneous model of DNA sequence evolution for phylogenetic analysis

N. Galtier, Manolo Gouy

Inferring Pattern and Process: Maximum-Likelihood Implementation of a Nonhomogeneous Model of DNA Sequence Evolution for Phylogenetic Analysis

Nicolas Galtier[1] *and Manolo Gouy*

Université Claude Bernard Lyon 1, Villeurbanne, France

A nonhomogeneous, nonstationary stochastic model of DNA sequence evolution allowing varying equilibrium G+C contents among lineages is devised in order to deal with sequences of unequal base compositions. A maximum-likelihood implementation of this model for phylogenetic analyses allows handling of a reasonable number of sequences. The relevance of the model and the accuracy of parameter estimates are theoretically and empirically assessed, using real or simulated data sets. Overall, a significant amount of information about past evolutionary modes can be extracted from DNA sequences, suggesting that process (rates of distinct kinds of nucleotide substitutions) and pattern (the evolutionary tree) can be simultaneously inferred. G+C contents at ancestral nodes are quite accurately estimated. The new method appears to be useful for phylogenetic reconstruction when base composition varies among compared sequences. It may also be suitable for molecular evolution studies.

Introduction

Many features of molecular sequence data make them suitable for statistical modeling. Characters are numerous and of a similar nature, and their evolution is partly constrained by global forces, applying at the gene or genome level (e.g., Sueoka 1961; Wolfe, Li, and Sharp 1987; Bernardi 1993; Jermiin et al. 1994). Molecular characters—i.e., sites in a multiple sequence alignment—can therefore be considered individuals of a population or outcomes of a random variable. Thus, phylogenetic reconstruction is seen as a statistical estimation problem, requiring models of DNA sequence evolution (Felsenstein 1988; Goldman 1990; Yang 1996). It is no surprise that model-based maximum-likelihood (ML) methods for phylogenetics have become increasingly popular as the use of molecular data has increased (Felsenstein 1981; Saitou 1988; Kishino and Hasegawa 1989; Goldman 1990; Yang 1993, 1996), although such methods were developed earlier (Cavalli-Sforza and Edwards 1967; Felsenstein 1973*a*, 1973*b*).

In stochastic Markov models of nucleotide substitution, rates of each kind of substitution (A→C, A→G, . . .) per time unit are given as functions of parameters of the model. This rate matrix represents the substitution process assumed. Usually, constancy of the rate matrix over the tree is assumed—the so-called homogeneity hypothesis. This means that in any lineage, DNA sequences converge toward a common equilibrium base composition (A, C, G, and T contents). Furthermore, constancy of base composition over the whole tree—the stationarity hypothesis—is also generally assumed: the assumed base composition in the ancestral sequence is equal to the equilibrium base composition of the assumed rate matrix and remains unchanged. If the homogeneity and stationarity assumptions were true, equal nucleotide frequencies would be expected in present-day sequences.

Actually, this quality of nucleotide frequencies is not the case in many data sets (e.g., Lockhart et al. 1992, 1994; Galtier and Gouy 1995): compositional changes are a major feature of genome evolution, making two common assumptions of substitution models unrealistic. Lockhart et al. (1994) and Galtier and Gouy (1995) showed that this departure from the model assumptions can mislead the standard tree-making methods: sequences of similar base composition are grouped, whatever their actual phylogenetic relationships. Alternative distance-based methods were devised to take into account unequal base compositions among sequences (Steel 1993; Lake 1994; Galtier and Gouy 1995). These new methods outperformed the usual ones when compositional biases were high.

Despite their practical success in phylogenetic analyses, the new methods do not increase knowledge about former evolutionary modes. They mainly extract the phylogenetic signal from the data regardless of peculiarities of base composition drift. Dealing with sequence pairs makes it difficult to recover information about the circumstances of past compositional changes. Yang and Roberts (1995) devised an ML implementation of a parameter-rich model accounting for both rate heterogeneity among sites and process heterogeneity among lineages, by assigning each branch in the tree its own model of substitution. However, this method was not tractable for more than four or five sequences. In this paper, we present an ML method for phylogenetic inference based on a new nonhomogeneous, nonstationary model. A reasonable number of sequences can be handled. The reliability of parameter estimates is assessed. The new method may be useful for recovering phylogenetic trees and/or studying molecular evolutionary processes.

Methods

The Model

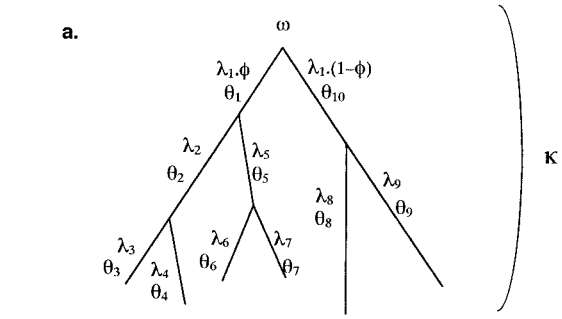
The evolutionary stochastic model we used was built to account for two of the major known forces gov-

[1] Present address: Laboratoire ''Génomes et Populations,'' Université des Sciences et Techniques du Languedoc, Montpellier, France.

Key words: molecular phylogeny, maximum likelihood, nonhomogeneous model, G+C content.

Address for correspondence and reprints: Manolo Gouy, CNRS UMR 5558, ''Biométrie, Génétique et Biologie des Populations,'' Université Claude Bernard Lyon 1, 43, Boulevard du 11 novembre 1918, 69622 Villeurbanne cedex, France. E-mail: mgouy@biomserv.univ-lyon1.fr.

Mol. Biol. Evol. 15(7):871–879. 1998



b.

parameters	symbol	number
ancestral G+C %	ω	1
branch lengths	λ_i	2*n* - 3
root location	φ	1
Ts/Tv ratio	κ	1
equilibrium G+C %	θ_i	2*n* - 2
		4*n* - 2

c.

	A	T	C	G
A	$-r_A$	$\frac{1-\theta}{2}$	$\frac{1-\theta}{2}$	$\kappa.\frac{1-\theta}{2}$
T	$\frac{1-\theta}{2}$	$-r_T$	$\kappa.\frac{1-\theta}{2}$	$\frac{1-\theta}{2}$
C	$\frac{\theta}{2}$	$\kappa.\frac{\theta}{2}$	$-r_C$	$\frac{\theta}{2}$
G	$\kappa.\frac{\theta}{2}$	$\frac{\theta}{2}$	$\frac{\theta}{2}$	$-r_G$

FIG. 1.—The evolutionary model used in this study in a six-species case. n = number of species. *a* and *b*, Parameter identification. *c*, Tamura's (1992) substitution rate matrix. r_A (respectively, r_T, r_C, r_G) is the sum of the substitution rates in the A (respectively, T, C, G) column: $r_A = r_T = (1 + \kappa\cdot\theta)/2$, $r_C = r_G = [1 + \kappa\cdot(1 - \theta)]/2$.

erning DNA evolution, namely, unequal transition (Ts) and transversion (Tv) rates and varying G+C content. One length λ_i is assigned to each branch of the unrooted topology. Tamura's (1992) model is used to represent the substitution process in each branch (fig. 1); this is a special case of the HKY model (Hasegawa, Kishino and Yano 1985). Tamura's model contains two parameters: Ts/Tv ratio (κ) and equilibrium G+C content (θ). In the present model, κ is kept constant all over the tree, but θ is allowed to vary from branch to branch. The model is neither homogeneous nor stationary, since equilibrium G+C content and expected base composition can vary among lineages. The assumed evolutionary process therefore lacks reversibility. This makes the likelihood dependent on the location of the root, which must be stated by the model (Yang and Roberts 1995); Felsenstein's (1981) ''pulley principle'' no longer applies. The precise location of the root in its branch is specified by a parameter, i.e., the fraction of the root branch length lying on the left side of the root (φ). The G+C content of the ancestral sequence is also a parameter of the model (ω), since the stationarity assumption does not apply.

The whole model includes five kinds of parameters (fig. 1): (1) ancestral G+C content, (2) location of the root in its branch, (3) Ts/Tv ratio, (4) branch lengths, and (5) equilibrium G+C contents in each branch. The assumed evolutionary process follows Tamura's (1992) model with variable θ among branches. In previous ML implementations of Tamura's model, θ was kept fixed over the tree. This model expands Galtier and Gouy's (1995) two-sequence model. For n species, the total number of parameters is $4n - 2$. Equal rates among sites are assumed. The substitution probabilities along a branch are given in the appendix.

Maximum-Likelihood Estimation

The likelihood L of a given tree is computed according to Felsenstein (1981), multiplying substitution probabilities along connected branches and summing over all possible ancestral states. The estimation process has two parts:

1. For a given (rooted) topology, find the ML estimates of parameters and record the ML value.
2. Perform step 1 for all competing (rooted) topologies, and pick up the one maximizing the ML value.

Step 1 is achieved by a modified Newton-Raphson algorithm close to Felsenstein and Churchill's (1996). Two tree-searching algorithms were implemented (step 2 above): star-decomposition (Saitou 1988; Yang 1995), and global rearrangements (pruning–regrafting) from any starting tree (Felsenstein 1993). Our versions of these algorithms are adapted to the rooted case.

Reliability of Estimates

Three approaches were used to investigate the reliability of the estimates of parameters, given a topology. First, putative local maxima were sought by varying the initial conditions of the iterative process for a given data set. Second, the sampling variance-covariance matrix of the parameter estimates was calculated, it is the opposite of the inverse of the information matrix (i.e., the matrix of second and cross second derivatives of the log-likelihood with respect to pairs of parameters) computed at the ML point (Edwards 1972). Third, the reliability of estimates was assessed by simulating DNA sequence evolution according to model assumptions and then applying the inference algorithm to the simulated data sets. Estimated parameters were then compared to actual ones. This operation may appear unnecessary, since the ML estimates of substitution parameters are known to be consistent, given a phylogeny (Edwards 1996). However, asymptotic properties tell us nothing about the reliability of estimates when a limited amount of data is used (Hillis and Huelsenbeck 1996). Furthermore, parameter estimation is achieved using an approximate iterative algorithm whose results are worth checking.

Results

Convergence

Contrasting with Yang and Roberts' (1995) report, few convergence problems were faced when Felsenstein and Churchill's (1996) modified Newton-Raphson algorithm was used. A six-species data set was used to search for putative local maxima. Small-subunit (SSU) rRNA sequences of six bacterial species were aligned (1,259 sites): *Aquifex pyrophilus* (Apy), *Thermotoga maritima* (Tma), *Thermus thermophilus* (Tth), *Deinococcus radiodurans* (Dra), *Bacillus subtilis* (Bsu), and *Escherichia coli* (Eco). The (Apy(Tma((Tth Dra)(Bsu Eco)))) topology was assumed (Olsen, Woese, and Overbeek 1994). The modified Newton-Raphson method was

applied 25 times with varying initial parameter values. The maximum log-likelihoods ranged from −5,476.1145 to −5,476.0966. This is two orders of magnitude less than log-likelihood differences among trees commonly considered significant. Parameter estimates were very close between replicates: the highest relative difference was found for parameter ϕ and was lower than 5%, while relative differences between estimates of a given branch length never exceeded 0.5%. Therefore, convergence artifacts seem unlikely to seriously mislead phylogenetic inferences.

Simulations

Three distinct simulation processes were performed. In the first one (S_1), a 10-species rooted tree topology (Kuhner and Felsenstein 1994), and a set of branch lengths were randomly drawn. Branch lengths were scaled so that their sum was 1.5; the length of the pathway connecting the most distantly related sequence pair on such trees averaged 0.75. No molecular clock was assumed. An ancestral G+C content value and 18 branch-specific equilibrium G+C contents were randomly drawn from a uniform distribution over [0, 100%]. The Ts/Tv ratio was set to 2. A 500-nt-long ancestral sequence was randomly drawn according to its expected base composition (A = T and C = G were assumed). The diverging evolution of 10 sequences was simulated following the above model. This procedure was repeated 100 times. Simulations S_2 and S_2' were built to assess the reliability of the ω estimate. In S_2, the ancestral G+C content was drawn within [0, 25%]. Equilibrium G+C contents θ_i were set to 90%. A 10-species topology with total length 1.5 was randomly drawn assuming the molecular clock hypothesis, so that all present-day sequences were equally distant from the root. This procedure generated present-day sequences with similar medium G+C contents (around 40%), while the ancestor was GC-poor. An opposite S_2' evolutionary process was also simulated: ω was drawn within [45%, 55%], and θ_i values were set to 10%, leading to 10 GC-poor present-day sequences. Five S_2 data sets and five S_2' data sets were generated.

We examined the estimability of all five classes of parameters. Residual standard deviations σ_ϕ, σ_κ, and σ_ω were computed for parameters ϕ, κ, and ω:

$$\sigma_\phi = \sqrt{\frac{1}{p} \cdot \sum (\phi_{est} - \phi_{act})^2},$$

where p is the number of replicates, ϕ_{est} and ϕ_{act} are the estimated and actual ϕ values for a given replicate, and the summation is over all p replicates (and similarly for σ_κ and σ_ω). Residual standard deviations over 100 S_1 simulations were σ_ϕ = 25.66%, σ_κ = 0.1789, and σ_ω = 1.24%. The location of the root on its branch appears to be poorly estimated; absolute residuals higher than 0.25 are common. In contrast, the κ estimate is reasonably reliable, and the ancestral G+C content ω is quite accurately recovered. Estimated versus actual branch lengths are plotted (fig. 2*a*). These estimates were compared with those of program FASTDNAML, based on a homogeneous stationary model. Residuals in both analyses (absolute values) are subtracted and plotted versus actual branch lengths (fig. 2*b*). For many branches, the homogeneous residual is significantly higher than the nonhomogeneous one, while the opposite effect is weaker. Thus, branch length estimates can be biased by unequal base compositions if the latter are not taken into account. Estimated versus actual equilibrium G+C contents are plotted (fig. 3*a*). A small part of these parameters lay on the boundaries of the relevance zone (0 or 1). Overall, a reliable global picture of past evolutionary modes could be extracted by the above inference algorithm from 500-nt-long sequences, although some θ_i values appear to be poorly estimated. θ_i residuals are plotted versus the lengths of underlying branches (fig. 3*b*). Equilibrium G+C contents in long branches appear to be correctly estimated, while those of short branches are not; when few substitutions happen, the evolutionary process can hardly be recovered. Equilibrium G+C contents in branches connected to the root are a special case: they are dependent on knowledge about the location of the root (see below). When the estimated ϕ value is very different from the actual one, θ_i values in the root branch are poorly estimated, while they are reliably

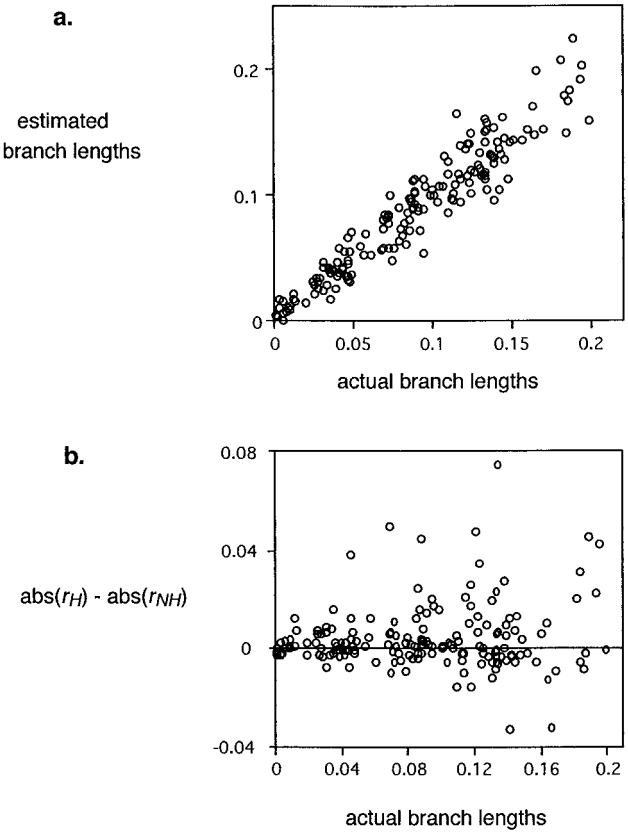


FIG. 2.—Branch length estimation from simulated S_1 data sets. *a*, Estimated versus actual branch lengths. *b*, Comparison between estimations based on a homogeneous model (program FASTDNAML; Olsen et al. 1994) and a nonhomogeneous model (this work), respectively. Residuals (estimated minus actual branch lengths) were computed in the homogeneous (r_H) and nonhomogeneous (r_{NH}) cases. abs = absolute value. Ten S_1 simulation repeats were randomly selected from 100 performed ones to make the figure clearer.

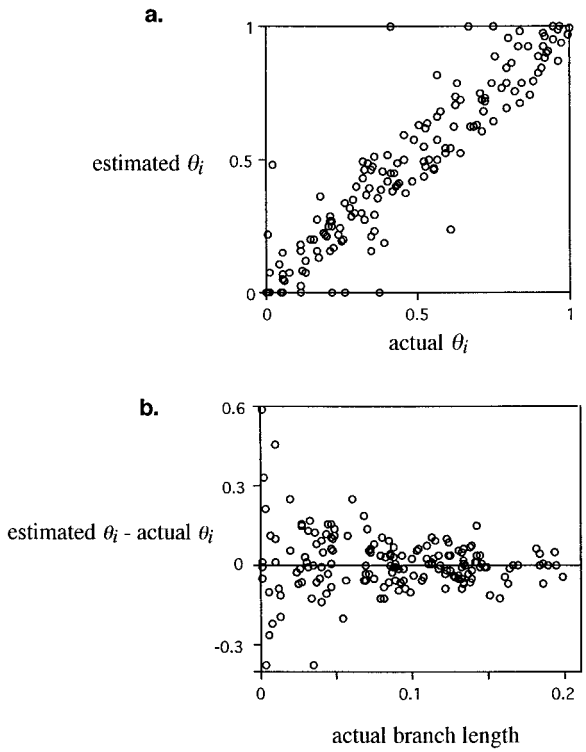


FIG. 3.—Equilibrium G+C content estimation from simulated S_1 data sets. *a*, Estimated versus actual θ_i. *b*, θ_i residual versus actual branch length. Ten simulation repeats are shown (see fig. 2 legend).

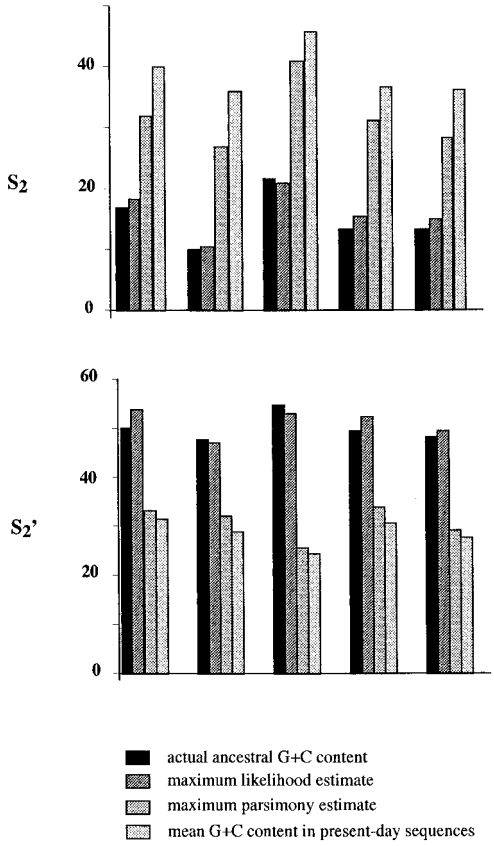


FIG. 4.—Ancestral G+C content estimation from five simulated S_2 data sets and five simulated S_2' data sets. The actual value and three estimates are shown. The maximum-parsimony estimate was computed by first recovering putative ancestral states at each site according to the maximum parsimony algorithm (Fitch 1971) and then calculating the fraction of G+C, accounting for equally parsimonious scenarios.

estimated if the actual ϕ value is given (not shown). The above results were obtained from trees with a length of 1.5 substitutions per site. Longer trees were also used (10 species, total lengths 2.0 and 3.0, resulting in average maximum pairwise distances 1.0 and 1.5, respectively), leading to similar results.

The accuracy of ancestral G+C content estimate is striking. However, since equilibrium G+C contents θ_i are randomly drawn, the mean G+C content in present-day sequences may be similar to the ancestral one for S_1 data sets. To test whether the above inference algorithm can actually extract information about ancestral base composition, simulations S_2 and S_2' were performed, in which the G+C contents of present-day sequences are nearly constant, but distinct from that of the ancestral one. Three ancestral G+C content estimates were computed from five S_2 data sets and five S_2' data sets: the mean of G+C contents of present-day sequences, parsimony site-by-site reconstruction of ancestral states, and our above algorithm. Results are shown in figure 4. As expected, the mean of present-day G+C contents is a poor estimate of ω. The parsimony-based estimate performs slightly better but is definitely not reliable. The nonhomogenous ML method gives quite accurate estimates for ω, even when G+C contents in present-day sequences are similar to each other and different from that of the ancestral one. This valuable property generalizes to the G+C content of any internal node of the tree, which can be deduced from the parameters of the model (not shown).

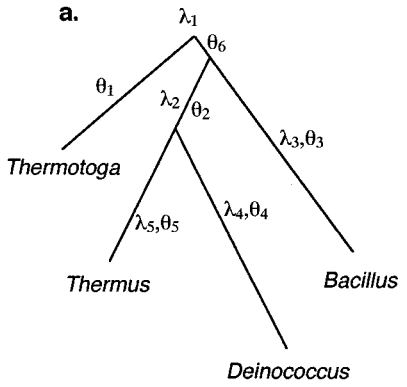
Six-species simulations were also conducted, and the same results were found (not shown). G+C contents at internal nodes were accurately recovered.

Sampling Variance and Covariance

The variance/covariance matrix was computed for real data sets of 4, 6, 8, and 10 species (bacterial SSU rRNA, 1,259 sites). A common pattern was found. Standard errors and correlation coefficients in the four-species case are given (fig. 5). Most results of the simulation process are confirmed: the standard error of parameter ϕ is high, those of κ, λ_i, and θ_i are reasonable, and that of ω is quite low. Further information is provided by the correlation coefficients of pairs of parameters. First, the location of the root ϕ and the equilibrium G+C contents in the branches connected to the root (θ_1 and θ_6 in fig. 5) are strongly correlated. No reliable information about the evolutionary mode within the root branch seems to be available from present-day sequences. Second, a systematic negative correlation between κ and branch lengths is found. In contrast, no significant correlation is found between the θ_i and λ_i values of a given branch, which is a valuable point.

Application

The new method was compared with some commonly-used ones using real data. A data set DS_1 was



b.

	estimate	s.e.
κ	2.073	0.201
φ	0.962	0.136
ω	61.8%	1.5%
λ_1	0.097	0.012
λ_2	0.042	0.008
λ_3	0.111	0.012
λ_4	0.118	0.013
λ_5	0.067	0.010
θ_1	81.4%	5,0%
θ_2	69.7%	8.2%
θ_3	28.6%	4.4%
θ_4	28.3%	4.2%
θ_5	71.4%	5.7%
θ_6	100%	15.1%

c.

	κ	φ	ω	λ_1	λ_2	λ_3	λ_4	λ_5	θ_1	θ_2	θ_3	θ_4	θ_5
φ	-19												
ω	-46	122											
λ_1	**-332**	-14	-94										
λ_2	-224	15	16	9									
λ_3	**-469**	-33	53	51	67								
λ_4	**-480**	-27	54	81	100	298							
λ_5	**-273**	65	-37	210	-43	14	-24						
θ_1	67	**-552**	-210	89	-8	-20	-43	-1					
θ_2	13	34	-3	-6	96	-6	32	-53	7				
θ_3	75	66	-33	4	-39	47	-85	59	22	48			
θ_4	74	35	-19	7	-56	-70	40	35	3	-93	67		
θ_5	60	55	-42	70	-32	-113	-123	210	18	173	81	116	
θ_6	30	**514**	-189	34	-17	-36	31	-82	-129	-54	-79	42	-62

FIG. 5.—Sampling standard errors and pairwise correlations between parameters in a four-species case (eubacterial SSU rRNA, 1,259 sites). *a,* Parameter identification. *b,* Parameter estimates and standard errors. *c,* Correlation coefficients for parameter pairs $\times 10^3$. Values higher than 0.25 or lower than −0.25 are shown in boldface. Correlation coefficients between the λ and θ parameters of a given branch are underlined.

built to check the accuracy of tree-making methods in the case of unequal base compositions. Small-subunit rRNA sequences of 16 bacterial species were selected from the rRNA database (Van de Peer et al. 1994). The species sampling was conducted this way: in seven bacterial phyla, two sequences were selected, namely, those with the highest and the lowest G+C contents. Phyla were defined according to Van de Peer et al. (1994). The monophyly of these groups is likely, since it is supported by many data sets (Woese 1987; Lloyd and Sharp 1993; Galtier and Gouy 1994; Eisen 1995). Two outgroup sequences (*Aquifex pyrophilus* and *Thermotoga maritima*) were added. Van de Peer et al.'s (1994) alignment was used. Ambiguously aligned regions were discarded (1,194 analyzed sites). Six tree-making methods were used: neighbor-joining (Saitou and Nei 1987) with Kimura's (1980) distance (NJ-K2), maximum parsimony (MP; Fitch 1971; Felsenstein 1993), ML based on a homogeneous model (ML-H; Olsen et al. 1994), NJ with the logdet distance (NJ-LD; Steel 1993; Lake 1994), NJ with Galtier and Gouy's (1995) distance (NJ-GG), and the new ML method based on a nonhomogeneous model (ML-NH). The latter three methods were devised to cope with unequal base compositions. For ML-NH, star decomposition was performed, and further rearrangements were tried manually. A limited number of trees could be examined due to extensive running time (star decomposition required 8 h on a Sun Sparc 1000). The number of correctly recovered phyla—i.e., the number of pairs of sequences of a given phylum actually grouped as neighbors—for each tree-making method are given in table 1. To assess the effect of unequal base compositions, a similar data set, DS_2, was constructed selecting sequences of nearly equal G+C contents within each phylum. Care was taken to ensure similar genetic distances between sequence pairs of a given phylum in both data sets. Table 1 shows that compositional biases are likely to mislead usual tree-making methods.

Table 1
Efficiencies of Six Tree-Making Methods in Recovering a Known Phylogeny from Bacterial SSU rRNA Sequences with Unequal (DS_1) or Nearly Equal (DS_2) G+C Contents

		Distance Methods			Likelihood Methods	
	MP	NJ-K2	NJ-LD	NJ-GG	ML-H	ML-NH
DS_1 . . .	1	0	3	5	3	5
DS_2 . . .	7	7	7	7	7	7

NOTE.—See text for discussion of the methods used.

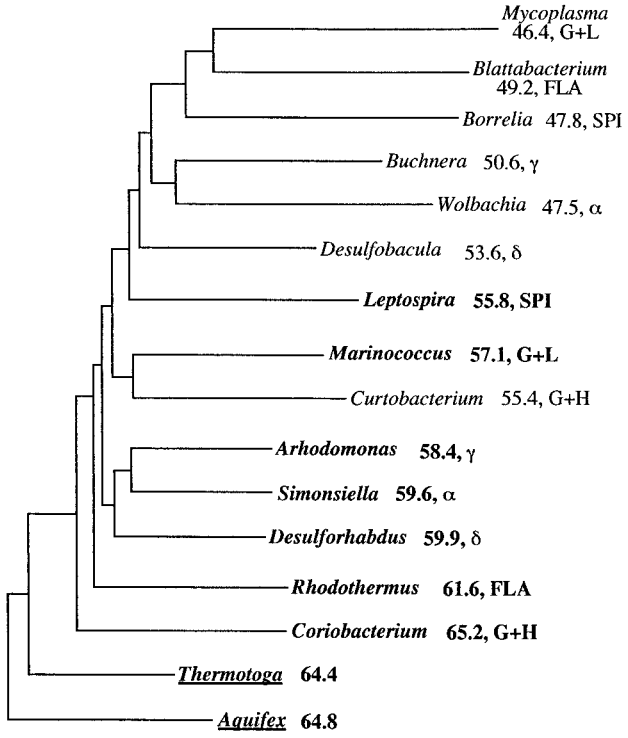


FIG. 6.—Neighbor-joining tree (Kimura's 1980 distance) for data set DS_1 (see text). Genus names, G+C contents, and phyla are given for 16 SSU rRNA eubacterial sequences. One genus per phylum (the G+C-rich one) is shown in boldface. Outgroups are underlined. Abbreviations: α, alpha proteobacteria; γ, gamma proteobacteria; δ, delta proteobacteria; FLA, *Flavobacterium* and relatives; G+H, high-GC Gram-positive bacteria; G+L, low-GC Gram-positive bacteria; SPI, Spirochetes.

This effect is striking for the NJ-K2 and MP methods: they recover all seven eubacterial phyla when sequences of nearly equal within-phylum base compositions are used, but zero or one when within-phylum G+C content variability is high. The NJ-K2 tree is given in figure 6: most G+C-rich sequences are ''attracted'' by the G+C-rich outgroups. The ML-H method looks more robust to the violation of the homogeneity assumption (see Galtier and Gouy 1995). Among distance-based methods, the NJ-LD and, especially, the NJ-GG methods appear more reliable than NJ-K2 when base compositions vary among sequences. Similarly, the new ML-NH method is more efficient than ML-H when the DS_1 data set is used.

Discussion

A new ML algorithm for phylogenetic inference is presented. It is based on a nonhomogeneous model of DNA evolution, allowing varying base compositions among sequences. Contrasting with Yang and Roberts' (1995) model, this model appears to be theoretically tractable for a large number of sequences, since no convergence problem was faced when the number of compared sequences increased. Much information about former evolutionary processes can be extracted from present-day sequences.

Relevance of the Model

The model implemented is parameter-rich, but most parameters are reliably estimated. Interestingly, the length λ_i and equilibrium G+C content θ_i of a given branch are not correlated. Representing equilibrium base frequencies in each branch by a single parameter (vs. three parameters in Yang and Roberts 1995) may be an improvement, since G and C contents (respectively, A and T contents) are nearly equal for long genomic fragments (Lobry 1995), except in mitochondrial genomes. Since the number of required parameters in a nonhomogeneous model is high, using Tamura's (1992) model rather than the more complex HKY one (in Yang and Roberts 1995) may be a valuable compromise, ensuring applicability. Some parameters, however, are poorly estimated. Among them, the location of the root on its branch and the equilibrium G+C contents in the root branch are highly correlated; little information about the evolutionary process in the root branch can be recovered. This may explain why the two-sequence approach failed in recovering former evolutionary processes (Galtier and Gouy 1995). Inaccurate estimates of G+C contents in short branches and negative correlations between the Ts/Tv ratio and branch lengths are two additional limitations of the present inference process. Both suggest the need to improve the parameterization of the model.

Regarding G+C contents in small branches, an excess of parameters is likely. In order to reduce this number, some θ_i's may be considered constant, or set to the mean of the θ_i's of connected branches. However, automatically choosing which θ_i's should be fixed this way is a difficult problem, because an equal number of parameters must be kept for all competing topologies if likelihoods are to be compared. Yang and Roberts (1995) suggested assuming a common θ_i value for a whole monophyletic group with similar base compositions, which is an interesting idea. However, the monophyly of the relevant group must be unchallenged throughout the analysis.

The negative correlation between the Ts/Tv ratio and branch lengths is probably the consequence of the implemented parameterization: equations (1) to (4) in the appendix show that parameter κ always multiplies λ_i. Using $\lambda_i' = \kappa \cdot \lambda_i$ as a new parameter is natural. However, this reparameterization does not work under the assumption of a common Ts/Tv ratio for all branches. Rejecting this assumption would probably remove the correlation between Ts/Tv ratios and branch lengths, but at a high cost, namely, $n - 3$ additional parameters, where n is the number of species. Since the correlation has no clear impact on the inference process, we favor our above model. An alternative solution would be an a priori estimation of the Ts/Tv ratio, possibly using a model allowing unequal rates among sites.

The present method correctly recovers the ancestral G+C content ω, while little information is extracted about the location of the root φ (and also the evolutionary process in the root branch), which may appear inconsistent. An interpretation of these results is as fol-

lows. Let r_1 and r_2 be the nodes connected to the root of the tree, and let ω_1 and ω_2 be the actual G+C contents at nodes r_1 and r_2. Estimation of ω_1 and ω_2 from the data is not dependent on knowledge about the location of the root. Indeed, our results show that these values are correctly estimated from usual data sets. When the ϕ parameter varies from 0 to 1, the optimal ω estimate should vary from ω_1 to ω_2. For a reasonably short root branch, ω_1 and ω_2 cannot differ much; so ω is correctly recovered even if ϕ is not known. If the root branch is very long, i.e., if the direction of evolution is not known in a large part of the tree, ω estimation may become less accurate.

Our main objective in this study was to check the relevance of a nonhomogeneous model; simplifying assumptions were made to focus on that point. Especially, equal rates among sites were assumed. This assumption has been found to be unacceptable for many data sets (Yang, Goldman, and Friday 1994; Tourasse and Gouy 1997). The Ts/Tv ratio becomes underestimated when unequal rates are not taken into account (Wakeley 1996). Since the above model appears tractable, generalizing it by removing the equal-rate assumption may be worthwhile.

Usefulness of the New Method

The present nonhomogeneous ML method may be useful for two distinct goals: (1) phylogenetic reconstruction and (2) study of the molecular evolution process.

Point 1 was exemplified by a 16-species study: the new method performed better than the usual ones, including ML implementations of homogeneous models and MP. Existing methods can become inconsistent when base composition varies among lineages—quite a common feature. Compositional biases have significant effects on phylogenetic reconstruction, as exemplified by data sets DS_1 and DS_2. Taking these effects into account greatly improves phylogenetic inferences.

The nonhomogeneous model may also be fitted to a data set, given a tree topology, to focus on former evolutionary processes. The sampling variances and covariances of inferred parameters can be used to assess their reliability. Ancestral G+C contents seem to be accurately estimated. This property may allow one to address a few long-standing questions of molecular evolution, including the evolution of isochores in vertebrate genomes (Mouchiroud, Gautier, and Bernardi 1988; Mouchiroud et al. 1991; Bernardi 1993) or the hypothesized G+C-richness of the ancestral genome of all living organisms (Woese 1987).

A major practical limitation of the new algorithm is running time. When the number of compared sequences is higher than seven or eight, only a small fraction of the tree space can be examined. Further, a priori knowledge about the root location is required, since rooted topologies are used. In view of this, the new method should be used to compare a reasonable number of phylogenetic hypotheses, possibly after an exhaustive preliminary analysis has been performed using faster methods. This strategy may be more efficient and faster than star decomposition, as suggested by empirical comparisons of several tree-searching algorithms (Z. Yang, personal communication).

Statistics and Systematics

The distribution of character states among species is the result of both evolutionary pattern (tree topology) and process (substitution rates). In our opinion, there is no reason why the process should be forgotten when inferring the pattern, as soon as a reasonable amount of information about it is available. Statistical modeling appears to be a suitable tool for inferring the history of molecular data. This is no surprise, since molecular characters are numerous and are undergoing global constraints: their states can be seen as outcomes of a random variable whose distribution derives from a stochastic process. A common criticism of model-based phylogenetic inferences is that models are not realistic. We would like to emphasize that even wrong or unrealistic models can lead to correct inferences, depending on the robustness of the estimates. Remarkably enough, the classical ML method based on a homogeneous model performed better than the MP method when base composition varied among sequences in our above example (see also Galtier and Gouy 1995); ML outperforms MP even when its underlying model is wrong. Presumably, this is because the MP method is not assumption free and is less robust than ML to departures from these assumptions (Yang 1996).

A computer program implementing the above method is available by anonymous ftp to: pbil.univ-lyon1.fr (pub/mol_phylogeny/nhml).

Acknowledgments

We thank Z. Yang for valuable advice about ML programming and one anonymous referee for excellent suggestions on an earlier version of this paper.

APPENDIX
Substitution Probabilities Under Tamura's 1992 Model

Let **R** be Tamura's (1992) matrix of instantaneous substitution rates (fig. 1), with parameters θ and κ. Suppose that such a process is undergone by a sequence evolving along a given branch of length λ. The matrix of substitution probabilities along the branch, **P**, is given by $\mathbf{P} = e^{\lambda \cdot \mathbf{R}}$. Explicitly:

$$p_{AA} = p_{TT} = \frac{1-\theta}{2} \cdot (1 + e^{-\lambda}) + \theta \cdot e^{-[(k+1)/2] \cdot \lambda} \quad (1)$$

$$p_{CC} = p_{GG} = \frac{\theta}{2} \cdot (1 + e^{-\lambda}) + (1-\theta) \cdot e^{-[(\kappa+1)/2] \cdot \lambda} \quad (2)$$

$$p_{GA} = p_{CT} = \frac{1-\theta}{2} \cdot (1 + e^{-\lambda}) - (1-\theta) \cdot e^{-[(\kappa+1)/2] \cdot \lambda} \quad (3)$$

$$p_{AG} = p_{TC} = \frac{\theta}{2}\cdot(1 + e^{-\lambda}) - \theta\cdot e^{-[(\kappa+1)/2]\cdot\lambda} \quad (4)$$

$$p_{AT} = p_{TA} = p_{GT} = p_{CA} = \frac{1-\theta}{2}\cdot(1 - e^{-\lambda}) \quad (5)$$

$$p_{CG} = p_{GC} = p_{TG} = p_{AC} = \frac{\theta}{2}\cdot(1 - e^{-\lambda}), \quad (6)$$

where p_{IJ} is the probability of state J at the bottom of the branch given state I at the top of the branch.

LITERATURE CITED

BERNARDI, G. 1993. The vertebrate genome: isochores and evolution. Mol. Biol. Evol. **10**:186–204.

CAVALLI-SFORZA, L. L., and A. W. EDWARDS. 1967. Phylogenetic analysis. Model and estimation procedures. Am. J. Hum. Genet. **19**:223.

EDWARDS, A. W. F. 1972. Likelihood. Cambridge University Press, Cambridge, England.

———. 1996. Assessing molecular phylogenies. Science **267**: 253.

EISEN, J. A. 1995. The recA protein as a model molecule for molecular systematic studies of bacteria: comparison of trees of recAs and 16S rRNAs from the same species. J. Mol. Evol. **41**:1105–1123.

FELSENSTEIN, J. 1973*a*. Maximum-likelihood estimation of evolutionary trees from continuous characters. Am. J. Hum. Genet. **25**:471–492.

———. 1973*b*. Maximum-likelihood and minimum-steps methods for estimating evolutionary trees from data on discrete characters. Syst. Zool. **22**:240–249.

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

———. 1988. Phylogenies from molecular sequences: inferences and reliability. Annu. Rev. Genet. **22**:521–565.

———. 1993. PHYLIP: phylogeny inference package. Version 3.5. Distributed by the author, Department of Genetics, University of Washington, Seattle.

FELSENSTEIN, J., and G. A. CHURCHILL. 1996. A hidden Markov model approach to variation among sites in rate of evolution. Mol. Biol. Evol. **13**:93–104.

FITCH, W. M. 1971. Toward defining the course of evolution: minimum change for a specific tree topology. Syst. Zool. **20**:406–416.

GALTIER, N., and M. GOUY. 1994. Eubacterial phylogeny: a new multiple-tree analysis method applied to 15 sequence data sets questions the monophyly of Gram-positive bacteria. Res. Microbiol. **145**:531–541.

———. 1995. Inferring phylogenies from sequences of unequal base compositions. Proc. Natl. Acad. Sci. USA **92**: 11317–11321.

GOLDMAN, N. 1990. Maximum likelihood inference of phylogenetic trees, with special reference to a Poisson process model of DNA substitution and to parsimony analyses. Syst. Zool. **39**:345–361.

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

HILLIS, D. M., and J. P. HUELSENBECK. 1996. Assessing molecular phylogenies. Science **267**:255–256.

JERMIIN, L. S., D. GRAUR, R. M. LOWE, and R. H. CROZIER. 1994. Analysis of directional mutation pressure and nucleotide content in mitochondrial cytochrome *b* genes. J. Mol. Evol. **39**:160–173.

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

KISHINO, H., and M. HASEGAWA. 1989. Evaluation of the maximum likelihood estimate of the evolutionary tree topologies from DNA sequences data, and the branching order in Hominoidae. J. Mol. Evol. **29**:170–179.

KUHNER, M. K., and J. FELSENSTEIN. 1994. A simulation of phylogeny algorithms under equal and unequal evolutionary rates. Mol. Biol. Evol. **11**:459–468.

LAKE, J. A. 1994. Reconstructing evolutionary trees from DNA and protein sequences: paralinear distances. Proc. Natl. Acad. Sci. USA **91**:1455–1459.

LLOYD, A. T., and P. M. SHARP. 1993. Evolution of the recA gene and the molecular phylogeny of Bacteria. J. Mol. Evol. **37**:399–407.

LOBRY, J. R. 1995. Properties of a general model of DNA evolution under no-strand-bias conditions. J. Mol. Evol. **40**: 326–330.

LOCKHART, P. J., 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.

LOCKHART, P. J., M. A. STEEL, M. D. HENDY, and D. PENNY. 1994. Recovering evolutionary trees under a more realistic model of sequence evolution. Mol. Biol. Evol. **11**:605–612.

MOUCHIROUD, D., G. D'ONOFRIO, B. AISSANI, G. MACAYA, C. GAUTIER, and G. BERNARDI. 1991. The distribution of genes in the human genome. Gene **100**:181–187.

MOUCHIROUD, D., C. GAUTIER, and G. BERNARDI. 1988. The compositional distribution of coding sequences and DNA molecules in humans and murids. J. Mol. Evol. **27**:311–320.

OLSEN, G. J., H. MATSUDA, R. HAGSTROM, and R. OVERBEEK. 1994. fastDNAml: a tool for construction of phylogenetic trees of DNA sequences using maximum likelihood. Comput. Appl. Biosci. **10**:41–48.

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.

SAITOU, N. 1988. Property and efficiency of the maximum likelihood method for molecular phylogeny. J. Mol. Evol. **27**: 261–273.

SAITOU, N., and M. NEI. 1987. The neighbor-joining method: a new method for reconstructing phylogenetic trees. Mol. Biol. Evol. **4**:406–425.

STEEL, M. A. 1993. Recovering a tree from the leaf colorations it generates under a Markov model. Appl. Math. Lett. **7**: 19–23.

SUEOKA, N. 1961. Variation and heterogeneity of base composition of deoxyribonucleic acids: a compilation of old and new data. J. Mol. Biol. **3**:31–40.

TAMURA, K. 1992. Estimation of the number of nucleotide substitutions when there are strong transitition-transversion and G+C-content biases. Mol. Biol. Evol. **9**:678–687.

TOURASSE, N. J., and M. GOUY. 1997. Evolutionary distances between nucleotide sequences based on the distribution of substitution rates among sites as estimated by parsimony. Mol. Biol. Evol. **14**:287–298.

VAN DE PEER, Y., I. VAN DEN BROECK, P. DE RIJK, and R. DE WACHTER. 1994. Database on the structure of small ribosomal subunit RNA. Nucleic Acids Res. **22**:3488–3494.

WAKELEY, J. 1996. The excess of transitions among nucleotide substitutions: new methods of estimating transition bias underscore its significance. TREE **11**:158–163.

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

WOLFE, K. H., W.-H. LI, and P. M. SHARP. 1987. Rates of nucleotide substitution vary greatly among plant mitochondrial, chloroplast, and nuclear DNAs. Proc. Natl. Acad. Sci. USA **84**:9054–9058.

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

———. 1995. Phylogenetic analysis by maximum likelihood (PAML). Version 1.1. Institute of Molecular Evolutionary Genetics, The Pennsylvania State University, University Park.

———. 1996. Phylogenetic analysis using parsimony and likelihood methods. J. Mol. Evol. **42**:294–307.

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

YANG, Z., and D. ROBERTS. 1995. On the use of nucleic acid sequences to infer branchings in the tree of life. Mol. Biol. Evol. **12**:451–458.

ROSS H. CROZIER, reviewing editor

Accepted March 24, 1998