Proc. Natl. Acad. Sci. USA
Vol. 92, pp. 11317–11321, November 1995
Evolution

Inferring phylogenies from DNA sequences of unequal base compositions

(molecular phylogeny/Markov model/G+C content)

NICOLAS GALTIER* AND MANOLO GOUY

Centre National de la Recherche Scientifique, Unité de Recherche Associée 2055, "Biométrie, Génétique et Biologie des Populations," Université Claude Bernard Lyon 1, 43, Boulevard du 11 novembre 1918, 69622 Villeurbanne cedex, France

Communicated by Carl R. Woese, University of Illinois at Urbana-Champaign, Urbana, IL, July 20, 1995

ABSTRACT **A new method for computing evolutionary distances between DNA sequences is proposed. Contrasting with classical methods, the underlying model does not assume that sequence base compositions (A, C, G, and T contents) are at equilibrium, thus allowing unequal base compositions among compared sequences. This makes the method more efficient than the usual ones in recovering phylogenetic trees from sequence data when base composition is heterogeneous within the data set, as we show by using both simulated and empirical data. When applied to small-subunit ribosomal RNA sequences from several prokaryotic or eukaryotic organisms, this method provides evidence for an early divergence of the microsporidian *Vairimorpha necatrix* in the eukaryotic lineage.**

Several distance-based algorithms of phylogenetic tree reconstruction are known to recover the true tree when the distance used is tree-like, i.e., when there is a tree and a set of branch lengths such that the distance between any sequence pair equals the length of the shortest path on the tree connecting the pair (1, 2). Obviously, the average number of nucleotide substitutions per site between two DNA sequences is such a tree-like distance. Therefore, properly estimating this value is a sufficient condition for reconstructing the actual phylogenetic tree. Estimating the average number of nucleotide substitutions per site between two DNA sequences involves making assumptions about the evolutionary process in both lineages. Departures of actual data from these assumptions are likely to lead to inaccurate distance estimates. Specifically, hypotheses of homogeneity (i.e., constancy of the process with time and among lineages) and stationarity (i.e., constancy of base composition within each lineage) are usually assumed. Both imply that nucleotide frequencies of all present-day sequences are equal, which is clearly wrong for many data sets: genomes of bacteria (3), animals (4), and plants (5) vary widely in their base composition. If sequences of unequal base compositions are compared, estimates obtained by the usual methods may become biased.

Effects of unequal base compositions on phylogenetic reconstructions have been little studied to date. Loomis and Smith (6) and Hasegawa *et al.* (7) suggested that ribosomal RNA-based phylogenetic trees may be misleading because of compositional bias. Saccone *et al.* (8) indicated that their evolutionary distance estimate became biased in the nonstationary case. Weisburg *et al.* (9) showed that trees derived from bacterial sequences with unequal nucleotide frequencies could be misleading and overcame the difficulty by a more stringent choice of sites. Steel *et al.* (10) proposed a modified version of Felsenstein's (11) parsimony-based test of phylogenies, taking base composition into account.

Here we show that violation of the homogeneity and stationarity hypotheses may strongly decrease the ability of tree-making methods in recovering the actual phylogenetic tree. We present an algorithm for estimating pairwise evolutionary distances without assuming homogeneity or stationarity of the evolutionary process. This distance estimate should be useful for phylogenetic analyses when compositional biases are observed.

THEORY

Two main factors are usually considered to bias nucleotide substitution rates: the transition/transversion ratio and the G+C content (12). We used an evolutionary model (Fig. 1) that was built to reflect these two factors. In this Markov model, the process in lineage 1 may differ from that in lineage 2, although both processes are constant in time. Sequence base composition may change with time in both lineages and need not be the same in present-day sequences 1 and 2. The total number of parameters is five: θ_0 stands for G+C content in the ancestral sequence, θ_1 and θ_2 stand for equilibrium G+C contents in lineages 1 and 2, α allows for unequal transition and transversion rates, and r is a rate parameter. This model reduces to Tamura's (13) three-parameter one when $\theta_1 = \theta_2 = \theta_0$, and to Kimura's (14) two-parameter one when the θ values are set to 0.5.

In this model, substitution rates per time unit from nucleotides A, C, G, and T in lineage i (i = 1 or 2) are given by

$$\lambda_{A_i} = \lambda_{Ti} = (1 + \theta_i\alpha)r/2$$

$$\lambda_{C_i} = \lambda_{G_i} = [1 + (1 - \theta_i)\alpha]r/2. \quad [1]$$

Let $A_i(t)$, $C_i(t)$, $G_i(t)$, and $T_i(t)$ be the base frequencies at time t in lineage i. Then K, the average number of nucleotide substitutions per site for T units of evolutionary time, is given by

$$K = \sum_{i=1}^{2}\int_0^T [A_i(t)\lambda_{A_i} + C_i(t)\lambda_{C_i} + G_i(t)\lambda_{G_i} + T_i(t)\lambda_{T_i}]dt. \quad [2]$$

Let us derive expressions of A, C, G, and T contents in lineage i. Obviously, at any date t,

$$C_i(t) = G_i(t) = \frac{1}{2} - T_i(t) = \frac{1}{2} - A_i(t). \quad [3]$$

The evolution of A_i is given by

$$A_i(t + dt) = A_i(t)(1 - \lambda_{A_i}dt) + \{[C_i(t) + \alpha G_i(t) + T_i(t)](1 - \theta_i)r\}dt/2. \quad [4]$$

The publication costs of this article were defrayed in part by page charge payment. This article must therefore be hereby marked "*advertisement*" in accordance with 18 U.S.C. §1734 solely to indicate this fact.

Abbreviations: SSU rRNA, small subunit ribosomal RNA; ML, maximum-likelihood; MP, maximum-parsimony; NJ, neighbor-joining; JC, Jukes and Cantor; TN, Tajima and Nei.

*To whom reprint requests should be addressed.

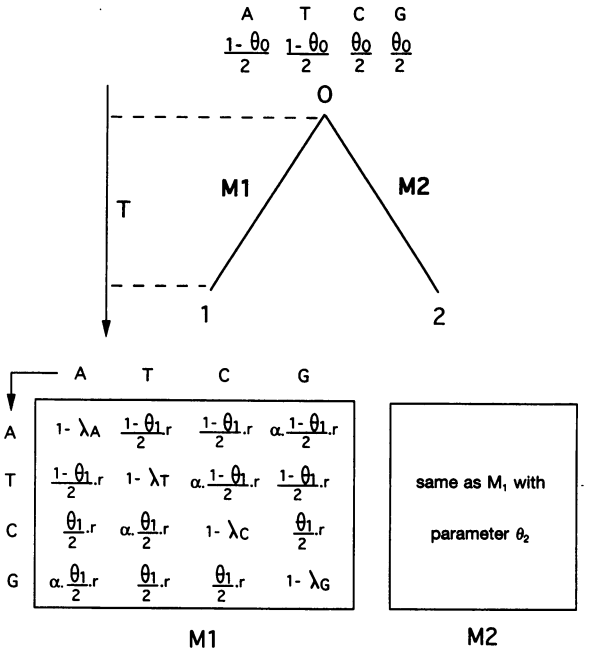


FIG. 1. Sequences 1 and 2 evolved from common ancestor 0 during time T. The ancestral base composition is described by G+C content θ_0. The A and T contents are assumed to be equal at time 0 (and will remain so in both lineages), as are the G and C contents. Rates of nucleotide substitutions are given by matrices M1 and M2, for lineages 1 and 2, respectively. Both processes follow Tamura's (13) model, but process 1 may not be the same as process 2. Parameters θ_1 and θ_2 are equilibrium G+C contents in lineages 1 and 2, respectively. λ_X is the substitution rate of nucleotide X in lineage 1.

Substituting Eqs. **1** and **3** into Eq. **4**, we obtain the differential Eq. **5**:

$$\frac{dA_i}{dt} = \lim_{dt \to 0} \frac{A_i(t+dt) - A_i(t)}{dt}$$

$$= -\frac{\alpha+1}{2} r A_i(t) + \frac{(1-\theta_i)(1+\alpha)}{4} r. \quad [5]$$

The solution of Eq. **5** with initial condition $A_i(0) = (1-\theta_0)/2$ is

$$A_i(t) = T_i(t) = \frac{1}{2} - C_i(t) = \frac{1}{2} - G_i(t)$$

$$= \frac{1-\theta_i}{2} + \frac{\theta_i - \theta_0}{2} e^{-(\alpha+1)rt/2}. \quad [6]$$

Substituting Eq. **6** into Eq. **3** and integrating the resulting equation leads to the expression of the evolutionary distance:

$$K = K_1 rT + K_2(1 - e^{-(\alpha+1)rT/2}) \quad [7]$$

where $K_1 = 1 + \alpha[\theta_1(1-\theta_1) + \theta_2(1-\theta_2)]$ and $K_2 = [\alpha/(\alpha+1)][(\theta_0 - \theta_1)(1-2\theta_1) + (\theta_0 - \theta_2)(1-2\theta_2)]$.

Estimates of parameters rT, θ_0, θ_1, θ_2, and α are required to compute K, the evolutionary distance.

Let Q be the observed proportion of sequence sites showing a transversion difference. Tamura (13) showed that the expected value of Q under his substitution model was independent of equilibrium frequencies θ_1 and θ_2. In Tamura's model, equalities $\theta_1 = \theta_2 = \theta_0$ were assumed, but these conditions are not necessary and the following equation indeed holds for the present model:

$$Q = (1 - e^{-2rT})/2. \quad [8]$$

Eq. **8** holds because in any lineage i, the pyrimidine-to-purine substitution rate is the sum of substitution rates from pyrimidine to A $[(1-\theta_i)r/2]$ and from pyrimidine to G $[\theta_i r/2]$, and similarly for purine-to-pyrimidine substitutions, so that the total transversion rate equals $2[(1-\theta_i)r/2 + (\theta_i r/2)] = r$, which does not depend on θ_i, the equilibrium G+C frequency. Parameter rT can therefore be estimated inverting Eq. **8**:

$$r\hat{T} = -\frac{1}{2}\ln(1-2Q). \quad [9]$$

It can be shown that under the present model parameter θ_0 is given by

$$\theta_0 = \frac{\theta_1 + \theta_2}{2} + \frac{1 - \theta_1 - \theta_2 + D}{2} e^{(\alpha+1)rT/2}, \quad [10]$$

where $D = y_{CC} + y_{CG} + y_{GC} + y_{GG} - y_{AA} - y_{AT} - y_{TA} - y_{TT}$, with y_{MN} being the observed proportion of sequence sites having nucleotide M in sequence 1 and nucleotide N in sequence 2.

We used approximate values for parameters θ_1, θ_2, and α, since providing exact analytical solutions happened to be a difficult task.

Estimates $\hat{\theta}_1$ and $\hat{\theta}_2$ of parameters θ_1 and θ_2 are given by the G+C contents of sequences 1 and 2, respectively. This assumes that the equilibrium base composition has been reached in both lineages. With this additional assumption, Eq. **10** reduces to

$$\hat{\theta}_0 = \frac{\hat{\theta}_1 + \hat{\theta}_2}{2}. \quad [11]$$

Finally, the transition/transversion ratio is assumed to be the same in all lineages. It is estimated once, from the whole data set. In the substitution model of Fig. 1, the ratio between the sums of transition and of transversion rates equals $\alpha/2$. This ratio is estimated following Kimura's (14) model for each sequence pair (i, j):

$$\frac{\alpha(i,j)}{2} = \frac{\ln[1 - 2P(i,j) - Q(i,j)] - \frac{1}{2}\ln[1 - 2Q(i,j)]}{\ln[1 - 2Q(i,j)]}, \quad [12]$$

where $P(i, j)$ is the observed proportion of sites in sequences i and j showing a transition difference, and $Q(i, j)$ is the observed proportion of sites showing a transversion difference. Estimate $\hat{\alpha}$ of parameter α is given by the mean of $\alpha(i, j)$ values for all sequence pairs. This estimate is used for all pairwise distance computations.

Substituting these five estimates into Eq. **7** provides an estimate $\hat{K}$ of the average number of nucleotide substitutions between the two sequences:

$$\hat{K} = -\frac{1}{2}\hat{K}_1 \ln(1-2Q) + \hat{K}_2[1 - (1-2Q)^{(\hat{\alpha}+1)/4}], \quad [13]$$

$$\text{with } \hat{K}_1 = 1 + \hat{\alpha}[\hat{\theta}_1(1-\hat{\theta}_1) + \hat{\theta}_2(1-\hat{\theta}_2)],$$

$$\text{and } \hat{K}_2 = \frac{\hat{\alpha}}{\hat{\alpha}+1}(\hat{\theta}_1 - \hat{\theta}_2)^2.$$

If we assume that sampling variances of estimates $\hat{\alpha}$, $\hat{\theta}_0$, $\hat{\theta}_1$, and $\hat{\theta}_2$ are very small, the sampling variance of $\hat{K}$ is

$$V(\hat{K}) = \left[\hat{K}_1 + \hat{K}_2 \frac{\hat{\alpha}+1}{2}(1-2Q)^{(\hat{\alpha}+1)/4}\right]^2 \frac{Q(1-Q)}{n(1-2Q)^2} \quad [14]$$

where $\hat{K}_1$ and $\hat{K}_2$ are defined as in Eq. **13** and n is the sequence length.

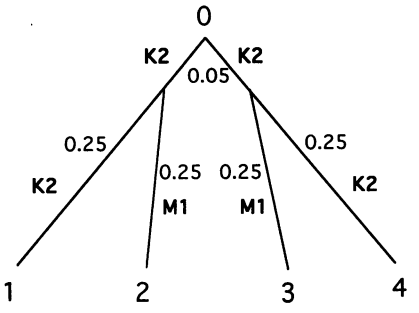


FIG. 2. Branch lengths are expected numbers of substitutions per site on the branch. Matrix M1 is given in Fig. 1, with θ_1 varying from 0.5 to 1 and α set to 2. Matrix K2 is that of Kimura's 1980 model (14), with the transition rate/transversion rate ratio set to 2.

RESULTS

Computer Simulations. A computer simulation was conducted to check the robustness of several tree-making methods in cases of unequal nucleotide frequencies. The model tree used is shown in Fig. 2. The ancestral sequence O was randomly drawn with equal probabilities for bases A, C, G, and T. Its length was 1000 bases. The evolutionary process in the whole tree followed Kimura's (14) two-parameter model with a transition/transversion ratio equal to 2, except for branches leading to taxa 2 and 3, where the evolutionary process followed a substitution scheme described by matrix M1 in Fig. 1 with $\alpha = 2$. The matrix M1, applied to both lineages 2 and 3, was varied between simulations by changing the value of parameter θ_1 from 0.5 to 1, thus providing sequences with increasing G+C contents. In all simulations, G+C contents of sequences 2 and 3 were roughly equal because they were derived from the same substitution matrix. For five tree-making methods, the proportion of trees in which the correct topology was recovered was plotted against the expected G+C-content difference ΔGC between sequence 2 (or 3) and sequence 1 (or 4). Constant (Fig. 3*a*) or gamma-distributed (Fig. 3*b*) substitution rates among sites were employed. The methods used were the maximum-parsimony (MP) method (15), the neighbor-joining (NJ) method (1) with Jukes and Cantor's (16) distance (NJ-JC), the NJ method with Tajima and Nei's (17) distance (NJ-TN), the maximum-likelihood (ML) method (18, 19), the NJ method with distances computed from transversions only (i.e., using Eq. **9**; NJ-eq9) and the NJ method with Eq. **13** distance (NJ-eq13).

The first three methods (MP, NJ-JC, and NJ-TN) are highly sensitive to compositional bias because their efficiency strongly decreases when ΔGC increases. In case of failure, the wrong topology supported is always tree [1,4][2,3], grouping together sequences with similar base compositions. MP, NJ-JC, and NJ-TN perform badly when ΔGC is higher than 10% (8% in case of unequal substitution rates among sites) and systematically fail when ΔGC exceeds 20%. Comparable results were obtained by using the NJ method with Kimura's (14) and Tamura's (13) distance estimates (data not shown).

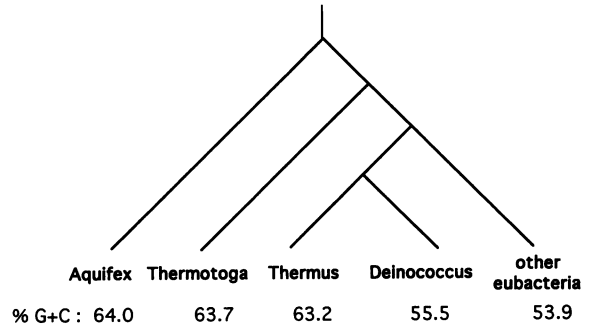


FIG. 4. "Other eubacteria" include *Chlamydia, Spirochaeta, Bacteroides, Agrobacterium, Escherichia, Fusobacterium, Clostridium, Bacillus, Micrococcus*, and *Anabaena*. % G+C is SSU-rRNA G+C content in positions retained for sequence comparisons. For other eubacteria, the average G+C content is given.

The ML, NJ-eq9, and NJ-eq13 methods are clearly more robust. Most importantly, NJ-eq13 performs better than ML and than NJ-eq9 for high ΔGC values with or without variable rates among sites. With variable rates among sites, a biologically more realistic situation, usual methods may become highly misleading for ΔGC values as low as 6% or 8%, depending on the shape and skewness of the gamma distribution.

Empirical Data. The efficiency of Eq. **13** for phylogenetic reconstructions was further questioned by using real data—namely, eubacterial small-subunit ribosomal RNA (SSU rRNA) sequences. Part of the phylogenetic tree of eubacteria is shown schematically in Fig. 4. The grouping of *Deinococcus radiodurans* and *Thermus thermophilus* in a single ancient phylum (9) and the earlier branchings of *Aquifex pyrophilus* and *Thermotoga maritima* (20, 21) are well-established features in the history of eubacteria. The SSU rRNA sequences of 14 eubacterial species, including *Aquifex, Thermotoga, Thermus*, and *Deinococcus* species, were employed to reconstruct phylogenetic trees by NJ-TN, ML, and NJ-eq13 methods. The G+C contents in these sequences are given in Fig. 4. *Aquifex, Thermotoga*, and *Thermus* are G+C-rich, whereas other bacterial species used, including *Deinococcus*, are not. Sequence alignment came from the Ribosomal Database Project (22), with further manual refinements. Only regions that could be aligned without ambiguity were used, and all positions with at least one gap were removed; 1235 sites were examined.

When all 14 species were used, NJ-eq13 and NJ-eq9 recovered the correct tree, whereas neither NJ-TN nor ML grouped G+C-rich *Thermus* with G+C-poor *Deinococcus*. The reli-

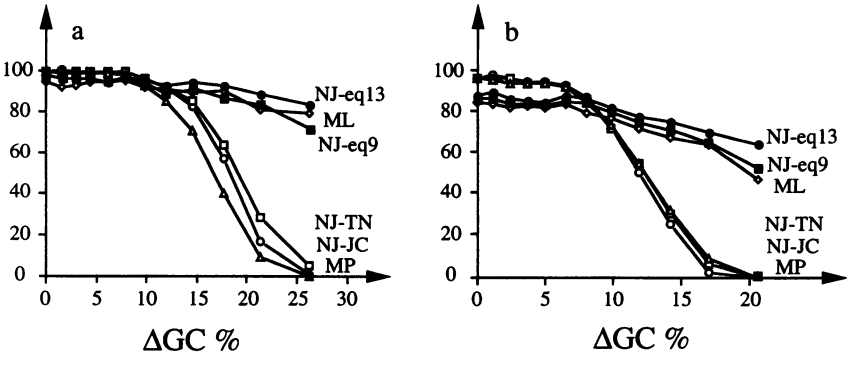
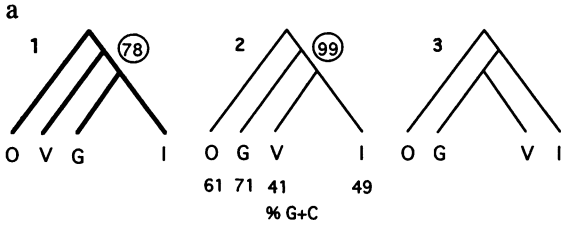


FIG. 3. Simulation was conducted according to Fig. 2. For each ΔGC value, 500 replications were performed. Lengths of sequences are 1000. (*a*) Equal substitution rates among sites. (*b*) Gamma distribution with shape parameter value 0.8 for substitution rates among sites.

abilities of the relevant internal branches were assessed by bootstrap analysis (23). The wrong *Aquifex/Thermotoga/Thermus* clade was supported by 985 bootstrap replicates out of 1000 when NJ-TN was used. This result emphasizes a well-known property of bootstrapping: if the tree-making method used is inaccurate, a wrong internal branch may be highly supported by bootstrap analysis. Thus, even well-supported internal branches may be wrong in case of compositional bias. The bootstrap score of the true tree recovered by NJ-eq13 was 93%. Five-species trees were also reconstructed by using *Aquifex, Thermotoga, Thermus, Deinococcus*, and one of the 10 remaining species. Consistently, NJ-eq13 and NJ-eq9 performed better than other methods, supporting the true tree in all 10 cases, whereas ML had 3 successes and NJ-TN none.

Eukaryote Phylogeny. Early branching orders in the eukaryotic domain were studied by using SSU rRNA sequences. Mitochondria-lacking diplomonads and microsporidia are considered the most ancient eukaryotic phyla, since many prokaryotic-like features were discovered in the primary and secondary structures of their SSU rRNA (24, 25). Furthermore, it is commonly admitted that diplomonads constitute the most ancient phylum (tree 2 in Fig. 5*a*), as that branching order is supported by usual tree-making methods (24, 26, 27). However, G+C-content differences between SSU rRNA of these species are extreme, as shown in Fig. 5*a*, suggesting that the usual methods may be misleading. We reexamined those data by using Eq. **13** to compute evolutionary distances.

Twelve archaeal and 14 eukaryotic SSU rRNA sequences were used, including those of the diplomonad *Giardia lamblia* and the microsporidian *Vairimorpha necatrix* (903 sites were used; alignments were from the Ribosomal Database Project). Archaea rather than Eubacteria were chosen as an outgroup because the similarity between archaeal and eukaryotic rRNA sequences is greater than that between eubacterial and eukaryotic sequences. As expected, the NJ-TN and ML methods reconstructed tree 2, but NJ-eq13 favored tree 1. Bootstrap percent values for the NJ-TN and NJ-eq13 methods are given in Fig. 5*a*. Four-species trees were also constructed, using all possible sequence quadruplets that include *Giardia* and *Vairimorpha* together with one archaeal and one eukaryotic sequence. Results are shown in Fig. 5*b*.



b

	NJ-TN	NJ-Eq13
tree 1	11	130
tree 2	133	14

FIG. 5. (*a*) Three alternative phylogenetic trees. G, *Giardia lamblia*; V, *Vairimorpha necatrix*; *O*, archaeal outgroup (*Halococcus, Halobacterium, Methanobacterium, Methanococcus, Methanothrix, Thermoplasma, Archaeoglobus, Thermococcus, Thermoproteus, Pyrodictium, Sulfolobus*, and *Desulfurococcus*); I, eukaryotic ingroup (*Crithidia, Physarum, Euglena, Dictyostelium, Entamoeba, Saccharomyces, Babesia, Glycine, Oryza, Drosophila, Xenopus*, and *Homo*). % G+C is SSU-rRNA G+C content in positions retained for sequence comparisons. For outgroup and ingroup, average G+C contents are given. Circled numbers are bootstrap percentage supports from 1000 replicates for the NJ-eq13 (tree 1) and NJ-TN (tree 2) methods. Bootstrap analysis was conducted with a reduced 12-species data set. Species used were *Halococcus, Methanobacterium, Thermoplasma, Thermoproteus, Sulfolobus, Giardia, Vairimorpha, Crithidia, Physarum, Dictyostelium, Saccharomyces*, and *Homo*. (*b*) Numbers of reconstructed trees of each alternative according to two distance computation methods for all 144 taxon quadruplets including *Giardia, Vairimorpha*, one outgroup, and one ingroup. Tree 3 was never reconstructed.

These results strongly suggest that the G+C-rich *Giardia lamblia* sequence is artifactually attracted by the relatively G+C-rich prokaryotic sequences when usual tree-making methods are used, and that tree 1 is the correct tree. Microsporidia may thus be the most ancient known eukaryotic phylum. Interestingly, quadruplets predicting tree 1 when the NJ-TN method is used are those involving outgroups of low G+C content and eukaryotic ingroups of high G+C content—i.e., cases where compositional effects are expected to be lower.

The NJ-eq9 method also supported the Archaea/Microsporidia grouping, but the overall branching order of the tree is highly dubious: mitochondrial *Physarum polycephalum* branches off deeper than amitochondrial *Giardia lamblia*.

Note that tree 2, which may not be the true tree according to the results above, gets a 99% bootstrap support when the NJ-TN method is used. Again, a dubious internal node is highly supported by bootstrap analysis when compositional effects are not taken into account.

DISCUSSION

Compositional bias among compared sequences can have strong consequences on phylogenetic studies: usual methods tend to group together sequences with similar base compositions, whatever their phylogenetic relationships. These wrong clades may be highly supported by bootstrapping.

The method of distance estimation presented here attempts to deal with varying compositional biases between species by using a Markov model of nucleotide substitutions. The present model, including initial conditions and distinct substitution processes in two diverging lineages, is more realistic than the usual models, whatever their particular substitution schemes, in cases of compositional bias: a major feature of evolutionary processes, diverging base compositions, is altogether ignored by the usual models. The present model also provides a theoretical basis for the use of transversions only rather than both transitions and transversions in cases of compositional bias: Eq. **9** shows that the transversion rate r does not depend on base frequencies. Further, Eq. **13** gives an estimate of the overall substitution rate by taking into account compositional biases, extracting more information from the sequences than the more basic transversion rate (Eq. **9**).

The distance estimated by our method provides clear improvements for phylogenetic tree reconstruction when sequences with unequal base compositions are compared, as both computer simulations and empirical data testing showed. When the stationarity and homogeneity hypotheses are assumed in evolutionary distance computation, the NJ method performs badly in cases of unequal base compositions, whereas it is quite efficient when Eq. **13** is used. The ML method is also robust in our computer simulations with constant rates among sites but performs badly when empirical data are used, as it fails to recover the true tree when 14 eubacterial sequences are employed and it fails in more than half of the cases when sequence quintuplets are used. Sensitivity to unequal rates among sites may be the reason for these failures, as suggested by computer simulations (Fig. 3*b*). Finally, the MP method is quite misleading in cases of compositional bias.

Thus, Eq. **13** should be useful for phylogenetic studies when there are strong differences between base compositions of sequences. This happens frequently in nature, especially when sequences from distantly related species are compared. Genomic G+C content varies widely between early-branching prokaryotic or eukaryotic species (see examples above), indicating that many different evolutionary processes were fol-

lowed. Clearly, we must take into account these compositional effects for a better understanding of deep phylogenetic events.

A computer program for the distance computation described above is available on request. This work was supported by the Groupe de Recherche "Informatique et génomes" of Centre National de la Recherche Scientifique.

1. Saitou, N. & Nei, M. (1987) *Mol. Biol. Evol.* **4,** 406–425.
2. Studier, J. A. & Keppler, K. J. (1988) *Mol. Biol. Evol.* **5,** 729–731.
3. Jukes, T. H. & Bushan, V. (1986) *J. Mol. Evol.* **24,** 39–44.
4. Bernardi, G. (1993) *Mol. Biol. Evol.* **10,** 186–204.
5. Montero, L. M., Salinas, J., Matassi, G. & Bernardi, G. (1990) *Nucleic Acids Res.* **18,** 1859–1867.
6. Loomis, W. F. & Smith, D. W. (1990) *Proc. Natl. Acad. Sci. USA* **87,** 9093–9097.
7. Hasegawa, M., Hashimoto, T., Adachi, J., Iwabe, N. & Miyata, T. (1993) *J. Mol. Evol.* **36,** 380–388.
8. Saccone, C., Pesole, G. & Preparata, G. (1989) *J. Mol. Evol.* **29,** 407–411.
9. Weisburg, W. G., Giovannoni, S. J. & Woese, C. R. (1989) *Syst. Appl. Microbiol.* **11,** 128–134.
10. Steel, M. A., Lockhart, P. J. & Penny, D. (1993) *Nature (London)* **364,** 440–442.
11. Felsenstein, J. (1985) *Syst. Zool.* **34,** 152–161.
12. Sueoka, N. (1988) *Proc. Natl. Acad. Sci. USA* **85,** 2653–2657.
13. Tamura, K. (1992) *Mol. Biol. Evol.* **9,** 678–687.
14. Kimura, M. (1980) *J. Mol. Evol.* **16,** 111–120.
15. Fitch, W. M. (1971) *Syst. Zool.* **20,** 406–416.
16. Jukes, T. H. & Cantor, C. R. (1969) in *Mammalian Protein Metabolism*, ed. Munro, H. N. (Academic, New York), pp. 121–123.
17. Tajima, F. & Nei, M. (1984) *Mol. Biol. Evol.* **1,** 269–285.
18. Felsenstein, J. (1981) *J. Mol. Evol.* **17,** 368–376.
19. Olsen, G. J., Matsuda, H., Hagstrom, R. & Overbeek, R. (1994) *Comput. Appl. Biosci.* **10,** 41–48.
20. Burgraff, S., Olsen, G. J., Stetter, K. O. & Woese, C. R. (1992) *Syst. Appl. Microbiol.* **15,** 875–878.
21. Woese, C. R. (1987) *Microbiol. Rev.* **51,** 221–271.
22. Olsen, J. G., Overbeek, R., Larsen, N., Marsh, T. L., McCaughey, M. J., Maciukenas, M. A., Kuan, W.-M., Macke, T. J., Xing, Y. & Woese, C. R. (1992) *Nucleic Acids Res.* **20,** 2199–2200.
23. Felsenstein, J. (1985) *Evolution* **39,** 783–791.
24. Sogin, M. L., Gunderson, J. H., Elwood, H. J., Alonso, R. A. & Peattie, D. A. (1989) *Science* **243,** 75–77.
25. Vossbrink, C. R., Maddox, J. V., Friedman, S., Debrunner-Vossbrink, B. A. & Woese, C. R. (1987) *Nature (London)* **326,** 411–414.
26. Shlegel, M. (1991) *Eur. J. Protistol.* **27,** 207–219.
27. Cavalier-Smith, T. (1993) *Microbiol. Rev.* **57,** 953–994.