15. S. Pitnick, G. S. Spicer, T. Markow, *Evolution* **51**, 833 (1997).
16. M. J. Bertram, D. M. Neubaum, M. F. Wolfner, *Insect Biochem. Mol. Biol.* **26**, 971 (1996); P. S. Chen, *Ann. Rev. Entomol.* **29**, 233 (1984); *Experientia* **52**, 503 (1996); L. G. Harshman and T. Prout, *Evolution* **48**, 758 (1993); L. A. Herndon and M. F. Wolfner, *Proc. Natl. Acad. Sci. U.S.A.* **92**, 10114 (1995); M. F. Wolfner, *Insect Biochem. Mol. Biol.* **27**, 179 (1997).
17. C. S. Price, *Nature* **388**, 663 (1997).
18. M. Aguadé, *Genetics* **150**, 1079 (1998); M. Aguadé, N. Miyashita, C. H. Langley, *ibid.* **132**, 755 (1992); A. Civetta and R. S. Singh, *J. Mol. Evol.* **41**, 1085 (1995); *Mol. Biol. Evol.* **15**, 901 (1998); S. C. Tsaur and C.-I. Wu, *ibid.* **14**, 544 (1997).
19. T. Prout and J. Bundgaard, *Genetics* **85**, 95 (1977).

20. We thank C. Langley for insightful discussion and J. Canale, A. Civetta, M. Dermitzakis, J. P. Masly, B. Todd, B. Wagstaff, B. West, and P. Whitley for assistance in scoring flies. This work was supported by an NSF grant to A.G.C. and an NIH grant to D.J.B.

25 August 1998; accepted 8 December 1998

A Nonhyperthermophilic Common Ancestor to Extant Life Forms

Nicolas Galtier,* Nicolas Tourasse, Manolo Gouy

The G+C nucleotide content of ribosomal RNA (rRNA) sequences is strongly correlated with the optimal growth temperature of prokaryotes. This property allows inference of the environmental temperature of the common ancestor to all life forms from knowledge of the G+C content of its rRNA sequences. A model of sequence evolution, assuming varying G+C content among lineages and unequal substitution rates among sites, was devised to estimate ancestral base compositions. This method was applied to rRNA sequences of various species representing the major lineages of life. The inferred G+C content of the common ancestor to extant life forms appears incompatible with survival at high temperature. This finding challenges a widely accepted hypothesis about the origin of life.

A remarkable feature of genomic sequences is their ability to retain traces of extremely ancient evolutionary events, including the very first steps of life on Earth. By sequencing small-subunit (SSU) rRNA genes from various eukaryotic and prokaryotic species in the late 1970s, Woese and colleagues could construct for the first time a comprehensive picture of the universal tree of life (*1*). This work gave rise to conjectures about the nature of the most recent common ancestor (MRCA) of extant life forms: A hot, auxotrophic origin of life was hypothesized (*1*). The information contained in molecular data, however, has been obscured by numerous base substitution events that occurred during thousands of millions of years of diverging evolution. Realistic modeling of the molecular evolutionary processes is required to discriminate between phylogenetic signal and noise (*2*). Here, we devised a Markov model accounting for three major forces governing DNA sequence evolution: unequal transition/transversion rates, unequal evolutionary rates among sequence sites, and varying G+C contents among lineages. Maximum likelihood inference based on this model applied to large-subunit (LSU) and SSU rRNA sequences yields insights about early molecular evolution. Our results cast doubts on one commonly accepted hypothesis, namely the thermophilic nature of the MRCA.

The designed Markov model of DNA sequence evolution generalizes Galtier and Gouy's nonhomogeneous model (*3*) by accounting for variable substitution rates among sites. In this model, the assumed substitution process in a given branch of the tree (that is, the probability of change from one nucleotide to another) depends on two parameters, namely transition/transversion ratio and equilibrium G+C content, that is, the G+C content that would be reached after infinitely long evolution (*4*). The latter parameter is allowed to vary between branches, so that G+C content can diverge with time and among lineages. This assumption appears necessary given the observed range of G+C content (40 to 75%) in actual rRNA sequences. The resulting model is nonhomogeneous (variable substitution process), nonstationary (equilibrium is not reached), and irreversible (*5*), in contrast to usual models of DNA sequence evolution. Substitution rates are highly variable among sites in rRNA molecules as a consequence of unequal selective constraints (*6*). Neglecting this point may lead to biased phylogenetic estimators (*7*). Following Yang (*8*), a discretized gamma distribution of rates was assumed to account for among-site rate variability.

In usual homogeneous-stationary models, the ancestral base composition of the compared sequences is deduced from the assumed substitution rate matrix. Here, ancestral G+C content is a free parameter that can be estimated by fitting the model to data. In a previous study, surprisingly accurate estimates of ancestral G+C contents were found from simulated data

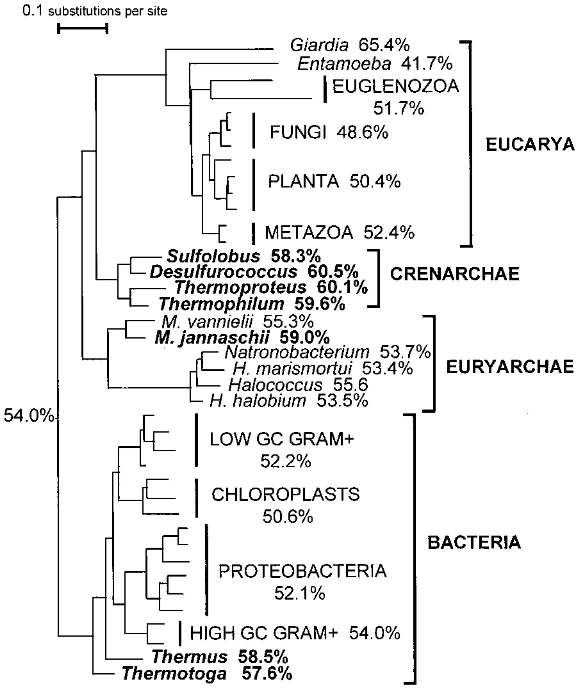

Fig. 1. Maximum likelihood tree reconstructed from 40 LSU rRNA sequences (*17*). G+C contents of (groups of) sequences are given next to taxon names. The inferred ancestral G+C content appears next to the root. Two hundred topologies obtained by rearranging an initial neighbor-joining tree were evaluated. The monophyly of Bacteria, Eucarya, Crenarchaea, Euryarchaea, Euglenozoa, animals, green plants, fungi, high-GC Gram-positive bacteria, low-GC Gram-positive bacteria, proteobacteria, and chloroplasts was kept in all trees, but the remaining branching orders were randomly shuffled; 1409 complete, unambiguously aligned sites were used. The names of thermophilic species are in boldface. *M*, *Methanococcus*; *H*, *Halobacterium*.

N. Galtier, Laboratoire de Biométrie, Génétique et Biologie des Populations, Université C. Bernard Lyon 1, France, and Laboratoire "Génome et Populations," Université Montpellier 2, 34095 Montpellier, France. N. Tourasse, Human Genetics Center, University of Texas, Houston, TX 77030, USA. M. Gouy, Laboratoire de Biométrie, Génétique et Biologie des Populations, Université C. Bernard Lyon 1, 69622 Villeurbanne, France.

*To whom correspondence should be addressed.

sets (*3*), suggesting that relevant information about former base compositions can be extracted from real data.

We estimated the G+C content in SSU and LSU rRNA of the MRCA by comparing sequences from 40 eukaryotic, bacterial, and archaebacterial species. For both molecules, a phylogenetic tree was reconstructed according to the distance-based neighbor-joining method (*9*) and rooted on the bacterial branch, as suggested by analyses of paralogous genes (*10*). The resulting tree topology was used to estimate the parameters of the above-described substitution model, including the ancestral G+C content. The estimated G+C content of the MRCA was 54.0% [95% confidence interval (c.i.) = (51.4, 56.5)] for LSU sequences and 56.1% [95% c.i. = (50.5, 60.4)] for SSU sequences. Confidence intervals were computed by parametric bootstrapping (*11*). These values are moderate: GC content varied from 47.2% to 65.4% (LSU) and from 43.7% to 70.4% (SSU) in the analyzed portion of compared sequences.

Additional analyses were performed on LSU data to assess the reliability of these estimates. First, sensitivity to the assumed topology was checked by reconducting the estimation procedure after modifying the model tree (Fig. 1). The ancestral LSU G+C content estimated from modified trees varied from 53.5% to 54.3%. Second, the ancestral LSU G+C content estimate was found to be robust to species sampling; it varied from 51.9% to 54.3%, depending on the set of species used (*12*). Third, a moderate (52.0%) ancestral LSU G+C content was still obtained after moving the root from the bacterial to the eukaryotic branch.

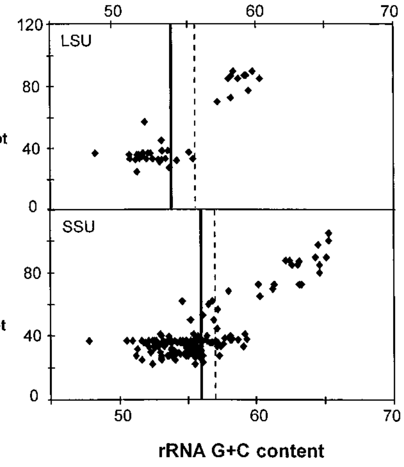

Fig. 2. Correlation between optimal growth temperature and rRNA G+C content in prokaryotes. Optimal growth temperatures (T_{opt}) were collected by Galtier and Lobry (*13*). G+C content was computed from those sites used in the present analysis. Top, LSU rRNA (41 species); bottom, SSU rRNA (177 species). Solid line, ancestral G+C content estimated from 40 sequences; dashed line, ancestral G+C content estimated from eight G+C-rich species. The shape parameter of the assumed gamma distribution (SSU, 0.60; LSU, 0.66) and the transition/transversion ratio (SSU, 3.25; LSU, 2.83) were estimated from the 40-species data sets and kept for the analysis of the eight-species subsets.

The rRNA ancestral G+C contents as estimated above are close to the average G+C contents in the compared sequences (LSU, 53.2%; SSU, 55.3%). We conducted an additional analysis by restricting the sampling to G+C-rich rRNA sequences. The G+C-richest two sequences of four domains (namely Eucarya, Bacteria, Crenarchaea, and Euryarchaea) were picked up, averaging to 60.2% G+C (LSU) and 66.1% (SSU). The ancestral G+C contents estimated from these extreme data sets were 55.5% [LSU, 95% c.i. = (53.4, 58.3)] and 57.3% [SSU, 95% c.i. = (51.9, 62.6)]—that is, less than the lowest G+C content in the compared eight sequences (57.4% and 62.4%, respectively). These striking results could hardly occur by chance. They strongly support the notion that the moderate ancestral G+C contents estimated from large data sets are not artifacts, but reflect actual information extracted from the data by the maximum likelihood method.

Prokaryotic rRNAs, together with tRNAs, are unique among nucleic acid sequences in that their G+C content is correlated with optimal growth temperature (T_{opt}): G-C pairs are more stable than A-U pairs at high temperatures because of an additional hydrogen bond (*13*). In contrast, the genomic G+C content of prokaryotes is not correlated to T_{opt}: The closed, double-stranded bacterial genome is not sensitive to high temperature. The correlation between rRNA G+C content and T_{opt} was reexamined by restricting the sequences to those sites used in the present analysis. The rRNA G+C content is significantly higher in hyperthermophilic prokaryotic species ($T_{opt} > 70$) than in mesophilic species (Fig. 2). The estimates of ancestral G+C contents (vertical lines) appear incompatible with a thermophilic life-style of the MRCA.

Thermophily is a widespread character found in many bacterial and archaeal phyla. Furthermore, the most deeply branching prokaryotes are thermophilic, and the branches leading to extant thermophilic species are generally short (Fig. 1). Woese (*1*) concluded from these observations that life had probably originated in a warm environment, consistent with what was commonly conjectured about the former temperature on Earth. This scenario was later endorsed by numerous authors (*14*) and reached the status of a working hypothesis for the early evolution of life. Forterre, however, has repeatedly criticized this view (*15*), arguing that present-day hyperthermophily may be a derived state.

Our maximum likelihood method yielded moderate estimates of the ancestral rRNA G+C contents, even from G+C-rich present-day sequences, although high G+C content in RNA is a necessary condition for survival in hot conditions. Following Forterre, we favor the notion that extant hyperthermophilic species evolved from mesophilic organisms via adaptation to high temperature. We argue that the short branches leading to thermophilic lineages do not reflect any affinity with the ancestor, but are the consequence of increased selective pressure acting on rRNA molecules of thermophilic species. The hypothesis of a hot origin of life cannot be ruled out (it may have preceded the MRCA), but no support from rRNA sequences can be claimed for it. Our results may be useful for dating the divergence between extant life forms if reliable knowledge about former temperatures on Earth becomes available.

References and Notes

1. C. R. Woese, *Microbiol. Rev.* **51**, 221 (1987).
2. In recent investigations of the age of the MRCA from protein data sets, estimates varied from 2000 to 6000 million years, exclusively depending on the assumed amino acid substitution model [R. F. Doolittle, D. F. Feng, S. Tsang, G. Cho, E. Little, *Science* **271**, 470 (1996); M. Hasegawa and W. M. Fitch, *ibid.* **274**, 1750 (1996); J. P. Gogarten, L. Olendzenski, E. Hilario, C. Simon, K. E. Holzinger, *ibid.*, p. 1750; X. Gu, *Mol. Biol. Evol.* **14**, 861 (1997)].
3. N. Galtier and M. Gouy, *Mol. Biol. Evol.* **15**, 871 (1998).
4. K. Tamura, *ibid.* **9**, 678 (1992).
5. Z. Yang and D. Roberts, *ibid.* **12**, 451 (1995).
6. A. Rzhetsky, *Genetics* **141**, 771 (1996).
7. N. Tourasse and M. Gouy, *Mol. Biol. Evol.* **14**, 287 (1997).
8. Z. Yang, *J. Mol. Evol.* **39**, 306 (1994). Eight equally probable classes of rates were used to approach the gamma distribution. All parameters were numerically estimated using the Newton-Raphson method. First and second derivatives of the likelihood function with respect to each parameter were analytically derived, excepting the shape parameter of the gamma distribution, whose derivatives were computed numerically.
9. N. Saitou and M. Nei, *Mol. Biol. Evol.* **4**, 406 (1987); N. Galtier and M. Gouy, *Proc. Natl. Acad. Sci. U.S.A.* **92**, 11317 (1995). The evolutionary distance used was a modified version of Galtier and Gouy's (1995) distance where among-site rate variation was taken into account assuming a truncated negative binomial distribution (*7*).
10. J. R. Brown and W. F. Doolittle, *Microbiol. Mol. Biol. Rev.* **61**, 456 (1997); S. L. Baldauf, J. D. Palmer, W. F. Doolittle, *Proc. Natl. Acad. Sci. U.S.A.* **93**, 7749 (1996).
11. Eighty data sets were generated by simulating the evolution of a randomly drawn DNA sequence along the tree of Fig. 1 using the maximally likely values of the parameters of the model. Ancestral G+C content was estimated for each data set. The highest two and lowest two values were removed. The remaining 76 ancestral G+C content values define a 95% confidence interval.
12. Two hundred data sets, each with 36 species, were built by randomly drawing representatives of the main eukaryotic, bacterial, and archaeal phyla among 167 available LSU rRNA sequences (*16*). For each data set, a neighbor-joining tree was reconstructed and ancestral G+C content estimated.
13. N. Galtier and J. R. Lobry, *J. Mol. Evol.* **44**, 632 (1997).
14. N. R. Pace, *Cell* **65**, 531 (1991); S. M. Barns, C. F. Delwiche, J. D. Palmer, N. R. Pace, *Proc. Natl. Acad. Sci. U.S.A.* **93**, 9188 (1996).
15. P. Forterre, *Cell* **85**, 789 (1996); *Curr. Opin. Genet. Dev.* **7**, 764 (1997).
16. P. De Rijk, Y. Van de Peer, R. De Wachter, *Nucleic Acids Res.* **24**, 92 (1996).
17. Data sets are available at http://pbil.univ-lyon1.fr/datasets/gcanc/.

23 July 1998; accepted 13 November 1998