

## THE EVOLUTION OF GENOMIC BASE COMPOSITION IN BACTERIA

ERIC HAYWOOD-FARMER AND SARAH P. OTTO<sup>1</sup>

Department of Zoology, University of British Columbia, Vancouver V6T 1Z4, Canada

<sup>1</sup>E-mail: otto@zoology.ubc.ca

**Abstract.**—Guanine plus cytosine (GC) content ranges broadly among bacterial genomes. In this study, we explore the use of a Brownian-motion model for the evolution of GC content over time. This model assumes that GC content varies over time in a continuous and homogeneous manner. Using this model and a maximum-likelihood approach, we analyzed the evolution of GC content across several bacterial phylogenies. Using three independent tests, we found that the observed divergence in GC content was consistent with a homogeneous Brownian-motion model. For example, similar rates of GC content evolution were inferred in several different bacterial subclades, indicating that there is relatively little rate heterogeneity in GC content evolution over broad evolutionary time scales. We thus argue that the homogeneous Brownian-motion model provides a good working model for GC content evolution. We then use this model to determine the overall rate of GC content evolution among eubacteria. We also determine the time frame over which GC content remains similar in related taxa, using a flexible definition for “similarity” in GC content so that, depending on the context, more or less stringent criteria may be applied. Our results have implications for models of sequence evolution, including those used for phylogenetic reconstruction and for inferring unusual changes in GC content.

**Key words.**—Ancestor states, maximum likelihood, bacteria, base composition, Brownian motion, GC content, sequence evolution.

Received December 12, 2001. Accepted February 18, 2003.

DNA base composition is one of the most straightforward genomic characteristics to measure and has been estimated in thousands of bacteria, where genomic guanine plus cytosine (GC) content ranges from 25% to 77% (Galtier and Lobry 1997). Although an extensive body of work explores the mechanisms that determine GC content in bacteria and discusses the evolutionary consequences for amino acid and codon preferences (e.g., Shields 1990; Gu et al. 1998; Pan et al. 1998; Chiusano et al. 1999; D’Onofrio et al. 1999), very little attention has been given to how this trait changes over time. This study aims to characterize the evolution of GC content in bacteria in terms of its rate, variability, and relationship to sequence evolution.

**GC content evolution.**—Two previous attempts have been made to characterize GC content evolution. Ochman and Lawrence (1996) noted that closely related bacteria are likely to have similar GC contents, that is, GC content displays a strong phylogenetic signal. As evidence, they presented a qualitative visual assessment of a tree of 17 bacterial species; in particular, they cited the compositional similarities among enteric bacteria as evidence for GC content stability. The small number of taxa and the qualitative appraisal method, however, limit the utility of this assessment. Gu et al. (1998) measured the correlation between divergence time and GC4 content (GC content of the groups of four codons specifying one amino acid) using Felsenstein’s (1985) independent contrasts test. The test found no correlation, implying that GC4 content is evolutionarily labile and exhibits little phylogenetic signal. This result contradicts the conclusion of Ochman and Lawrence (1996) but was based on a phylogeny spanning a much longer time frame. A quantitative analysis of genomic GC content over a broad range of evolutionary time frames is needed to evaluate the degree of phylogenetic signal in GC content.

How might GC content change over time? First consider a neutral model where the underlying transition rates remain

constant, with a mutation rate from AT to GC of  $\gamma\mu$  and a mutation rate from GC to AT of  $(1 - \gamma)\mu$ . Thus,  $\gamma$  measures the mutational bias toward GC and varies between zero and one. At steady state, the expected GC content under this model is equal to  $\gamma$ . If the mutational bias,  $\gamma$ , remained constant over time, it would be essentially impossible to explain the enormous variation in GC content observed among bacteria. For example, if  $\gamma$  were always 0.5 and if GC content were estimated from 30 kilobases, 99.9% of very distantly related bacteria would have GC values between 49% and 51%, which is much less than the observed range.

Sueoka (1993) has suggested that GC content evolves as a result of the emergence of new alleles at DNA replication and repair genes. This causes the relative frequencies of different types of mutations to shift, changing the mutational bias,  $\gamma$ , of a lineage. Because new alleles, including those responsible for changes in mutational pressures, arise as a function of the overall genomic mutation rate, which is remarkably constant among bacteria (Drake 1991), we might expect  $\gamma$  to change over time at a rate that is roughly similar across lineages of bacteria. If, in any period of time, the mutational bias of a lineage increases or decreases with equal probability, GC content would follow a random walk over time, with an expected GC content equal to its initial value, but with variation around this expectation. In reality, it is almost certainly the case that  $\gamma$  may change in discrete steps following mutations (including gene loss or gain) in the replication and repair machinery of a cell. The changes in GC content caused by changes in the mutational process ( $\gamma$ ) will, however, take time to accumulate. Furthermore, over a broad enough time scale, a random walk with discrete steps will behave similarly to a random walk following a continuous path.

Potentially, selection might also cause changes in GC content. Over short time scales, we would expect selective forces to be correlated over time, such that GC content would change

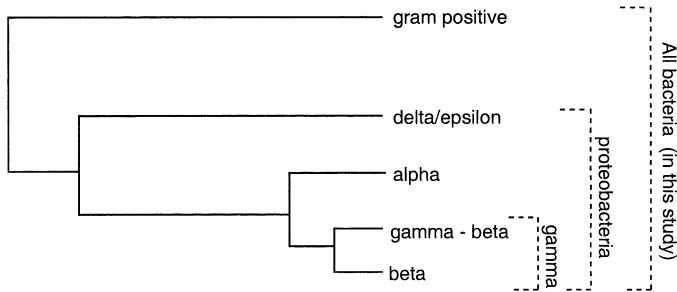


FIG. 1. Phylogenetic relationship among the bacterial groups used in this analysis.

in a consistent direction rather than follow a random walk. However, the direction of selection may be very weakly correlated over longer time scales. That is, GC content might be buffeted around by selection, sometimes increasing and sometimes decreasing. In this case, GC content would again approximate a random walk when examined over sufficiently long time periods.

In this paper, we explore the use of a Brownian-motion model to describe the evolution of GC content over time. A Brownian-motion model describes a random walk in continuous time over a continuous state space, where future states depend only on the current state of the system (Taylor and Karlin 1998). In mathematical terms, we assume that GC content follows a time-homogeneous diffusion process without drift. The main assumptions that we are invoking with this Brownian-motion model are that there is no general trend toward lower or higher GC content (specifically, the expected GC content over all species does not change over time) and that the rate of GC content evolution does not change among species or over time. A consequence of these assumptions (Taylor and Karlin 1998) is that GC content among descendants of an ancestor should follow a normal distribution with a mean equal to the ancestral GC content and with a variance that rises linearly with time.

To assess the appropriateness of the Brownian-motion model, we perform three tests, which ask: (1) Is genomic GC content divergence related to phylogenetic distance as expected under Brownian motion? (2) Is the variance around this expectation consistent with Brownian motion? and (3) Is the rate of change of genomic GC content the same in independent clades of bacteria? Tests (1) and (2) examine whether the observed pattern of divergence matches that expected from a Brownian-motion model, whereas test (3) examines whether the assumption that GC content evolves at roughly the same rate in different taxa is warranted (i.e., it tests the homogeneity assumption). On the basis of these tests, we fail to reject the Brownian-motion model. We then proceed to use this model to characterize GC content evolution. In particular, we estimate the overall rate at which GC content has evolved in a broad array of eubacterial species. Using this estimate, we then ask over what time period can we expect related bacteria to have similar genomic GC contents, using different criteria for "similarity." In this study, we use genomewide GC content information from the bacteria listed in Table 1 (see Fig. 1) and infer rates of change in GC content over time using the ancestor maximum like-

lihood (ANCML) program of Schluter et al. (1997; <http://www.zoology.ubc.ca/~schluter/ancml.html/>). Given GC content information and a phylogeny (inferred from 16S rDNA) for a group of species, we obtain maximum-likelihood estimates for the rate of GC content evolution and for ancestral GC values.

## METHODS

**Base composition.**—The genomic base compositions of 78 eubacterial species were obtained from the Codon Usage Database (Nakamura et al. 1997; <http://www.kazusa.or.jp/codon/>), which provides codon usage and GC content information based on the protein-coding genes available in GenBank (Table 1, Fig. 1). To ensure that the genomic estimates were accurate, species for which less than 30 kilobases were available were excluded from the analysis. In most comparisons, we focused on longer-term relationships, and only one species per genus was used (the one with the greatest number of base pairs sequenced). In an analysis of the shorter-term relationships within the Enterobacteriaceae group (taxa 22–29 and 69–78), all species were included that had sufficient data, that is, they were present in the Codon Usage Database with GC content estimated from more than 30 kilobases and in the 16S rDNA database, except if they were identical in 16S rDNA sequence to another enteric in our database. For species pairs with identical 16S rDNA sequences, evolutionary distance is inaccurately estimated to be zero, which causes ANCML to infer, erroneously, that there is an infinite rate of change in GC content if the species differ to any extent in genomic GC content.

**Phylogenetic reconstruction.**—Under a Brownian-motion model, we expect GC content to follow a random walk over time with a constant rate parameter, determined by the rate of change of mutational pressures (i.e., changes in  $\gamma$ ) and/or selective pressures. To determine this rate parameter, we would ideally relate GC content changes to the amount of geological time separating taxa, but the lack of a good fossil record for bacteria precludes this possibility. As a substitute for geological time, we use phylogenies inferred from 16S rDNA sequence data. 16S rDNA genes are sufficiently conserved that aligning DNA sequences and inferring phylogenetic relationships among diverse bacterial species is feasible (Ludwig et al. 1998). With respect to GC content, 16S rDNA genes exhibit much less variation in base composition among species than protein-coding genes, perhaps because all sites are nonsynonymous (in this study, the standard deviation across species in %GC was only 2.5 among 16S rDNA genes but 12.4 among genomes; Table 1). Any biases or errors in the phylogenetic reconstruction will, however, introduce errors in our estimates of GC content evolution, which is an unfortunate limitation to our study.

Trees were extracted from the Ribosomal Database Project's (RDP) Suggest Tree program (Maidak et al. 1997; <http://rdp.life.uiuc.edu/RDP/commands/sgtree.html>), which constructs trees using fastDNAML (Olsen et al. 1994), a maximum-likelihood method based on Felsenstein's (1981, 1993) DNAML program. Sequence alignments and maximum-likelihood parameters in RDP are determined by specialists, who take into account secondary structure, site-specific mutation

rates, and other features (Ludwig et al. 1998). All trees downloaded from RDP included the archaea *Methanococcus jannaschii* as a known outgroup to the bacteria studied; the outgroup was then removed manually to avoid biasing the ANCML rate calculations. The 16S rDNA phylogenies inferred by RDP do not assume a molecular clock, and their branch lengths are in terms of the number of nucleotide substitutions at sites whose rate of evolution is at the median of the distribution of mutation rates. Throughout this paper, divergence times and rate parameters are scaled by the branch lengths on the 16S rDNA phylogenies.

*Estimation of rates and ancestor states.*—Genomic GC content of ancestral nodes and rates of change were calculated using ANCML (Schluter et al. 1997; <http://www.zoology.ubc.ca/~schluter/ancml.html/>). This program uses a Brownian-motion model to approximate the change in a continuous trait over time. Given a phylogeny and the trait value (genomic GC content) of each extant species, ANCML uses a maximum-likelihood approach to calculate the rate,  $\beta$ , at which the variance of descendant traits increases over time and to estimate the ancestral states for all internal nodes in the phylogeny. (Note that  $\beta$  is estimated by integrating over all possible ancestral states at the internal nodes and hence does not rely on the accuracy of any particular ancestral reconstruction.) The expected squared difference between two species  $i$  and  $j$  with trait values  $\mu_i$  and  $\mu_j$  is, by definition,  $\beta$  multiplied by the divergence time ( $t$ ) since the descendants shared a common ancestor (Schluter et al. 1997). Alternatively, if time is measured as the sum of branch lengths separating two taxa ( $T_{ij} = 2t$ ),

$$E[(\mu_i - \mu_j)^2] = \beta T_{ij}/2. \quad (1)$$

It can be shown that equation (1) also describes the expected squared difference between an ancestor and a descendant,  $E[(\mu_i - \mu_a)^2]$ , with  $T_{ij}$  being replaced by  $T_{ai}$ , the sum of branch lengths separating the ancestor and descendant. Thus, given  $\beta$ , the standard deviation in trait values expected among descendants can be calculated as a function of the divergence time since their most recent common ancestor ( $T_{ai}$ ):

$$\begin{aligned} \text{SD}_{\text{descendants}} &= \sqrt{E\{[\mu_i - E(\mu_i)]^2\}} = \sqrt{E[(\mu_i - \mu_a)^2]} \\ &= (\beta T_{ai}/2)^{1/2}. \end{aligned} \quad (2)$$

If variation is instead measured as the absolute difference in GC content between two taxa, properties of a normal distribution can be used to show that:

$$E[|\mu_i - \mu_j|] = (\beta T_{ij}/\pi)^{1/2} \quad \text{and} \quad (3a)$$

$$E[|\mu_i - \mu_a|] = (\beta T_{ai}/\pi)^{1/2}. \quad (3b)$$

Letting  $\beta$  equal the rate estimated from data,  $\hat{\beta}$  equal the true value of this parameter, and  $N$  equal the number of ancestral nodes, the function  $\beta N/\hat{\beta}$  has a  $\chi^2$  distribution with  $N$  degrees of freedom (Schluter et al. 1997). Thus, 95% confidence intervals for  $\beta$  can be obtained from:

$$\beta N/\chi_{(N,0.975)}^2 < \hat{\beta} < \beta N/\chi_{(N,0.025)}^2. \quad (4)$$

#### ANALYSES AND RESULTS

To assess whether changes in GC content fit a Brownian-motion model, we examined GC content differences between

20 phylogenetically independent species pairs (Table 1). These species pairs were drawn from throughout the phylogenies shown in Figures 2a–e, with the restriction that no branch was included twice. These species pairs represent a wide range of divergence times (0.03–0.48 substitutions/site), tabulated by adding the length of all branches separating the two species on the 16S rDNA phylogeny ( $T_{ij}$ ). In Figure 3, we plot the absolute GC content difference against the divergence time for each species pair.

As a first test, we asked whether the absolute change in GC content rises as a square-root function of divergence time, as expected from equation (3a). Using the linear regression package in Mathematica (Wolfram 1991), GC differences ( $Y_i$ ) were fit to divergence times ( $X_i$ ) using the model:

$$Y_i = a + b(X_i)^{1/2}. \quad (5)$$

A  $t$ -test indicated that the best-fitting  $b$ -value was significantly different from zero ( $P < 0.005$ ) but that  $a$  was not ( $P = 0.186$ ), which is consistent with the expectation of the Brownian-motion model. Furthermore, the fit is not significantly improved if an additional, higher-order term,  $cX_i$ , is added to the model. The best-fitting curve of the form  $Y_i = b(X_i)^{1/2}$  is shown on Figure 3, with an estimated  $b = 31.97$ , which is equivalent to an estimated  $\beta = 3210$  (see eq. 3a).

As a second test, we note that the Brownian-motion model predicts that there should be substantial variance around the best-fitting square-root function. To test whether the observed proportion of the variance explained by the square-root model ( $R_{\text{obs}}^2 = 0.741$  for  $Y_i = 31.97[X_i]^{1/2}$ ) differed significantly from the expectation under a Brownian-motion model, we carried out a parametric bootstrap test. Using the observed divergence times and the inferred  $\beta$  value, we generated GC content differences between each pair of taxa by drawing random variates from a normal distribution with mean zero and variance given by equation (1). After generating GC differences between each pair of species, we found the best-fitting square-root function to the data (assuming  $a = 0$ ) and calculated  $R^2$ . This procedure was repeated 1000 times, and the bootstrap distribution for  $R^2$  was compared to  $R_{\text{obs}}^2$ . The observed 95% confidence interval for  $R^2$  based on the bootstrap replicates was 0.48 to 0.81. Thus, the fit of the observed data to a square-root function ( $R_{\text{obs}}^2 = 0.741$ ) was well within the distribution obtained by parametric bootstrapping (177/1000 of the simulated  $R^2$ -values were larger than  $R_{\text{obs}}^2$ ). Note that we would have rejected the Brownian-motion model if the fit were substantially worse ( $<0.48$ ) or better ( $>0.81$ ) than observed. Although we initially used a  $\beta$ -value of 3210, this test of the Brownian-motion model is not sensitive to the value of  $\beta$ . Indeed, similar distributions for  $R^2$  were obtained using  $\beta = 10^{-6}$  and  $\beta = 10^6$ .  $\beta$  affects the scale along the  $y$ -axis, which measures the amount of divergence between pairs of taxa, but  $\beta$  does not affect the amount of scatter expected around the best-fitting square-root function as measured by  $R^2$ , presumably because both the total variance and the variance explained by the model increase linearly with  $\beta$ . Thus, the shape of the best-fitting curve and the amount of scatter around this curve are both consistent with a Brownian-motion model.

As a third test of the homogeneous Brownian-motion model, we compared the rates of GC content evolution among

TABLE 1. Table of taxa used in this study. ID numbers identify taxa in Figure 2. An asterisk after the species name indicates that it was included in the analysis of a broad mixture of bacteria. A number in parentheses after the species name designates the 20 phylogenetically independent species pairs. RDP ID is the code used to identify sequences in the Ribosomal Database Project's Suggest Tree program (Maidak et al. 1997; <http://rdp.life.uiuc.edu/RDP/commands/sgtree.html>). Genomic GC content for protein-coding genes was obtained from the Codon Usage Database (Nakamura et al. 1997; [http://www.genome.ad.jp/kegg/catalog/org\\_list.html](http://www.genome.ad.jp/kegg/catalog/org_list.html); 15 August 2002 database). GC content for 16S rDNA genes was calculated from the sequences in RDP (with GenBank accession numbers where available).

	ID	Species	RDP ID	%GC (genomic)	%GC (16s rDNA)	GenBank Accession
Proteobacteria (see also 66–78)						
Alpha	1	<i>Gluconacetobacter xylinus</i> *	Aba.xylinm	63.21	56.33	X75619
	2	<i>Rickettsia prowazekii</i> * (20)	Ric.prowaz	30.63	50.53	M21789
	3	<i>Rhodobacter capsulatus</i> (20)	Rb.capsula	66.34	54.73	M60671
	4	<i>Zymomonas mobilis</i> * (2)	Zym.mobili	47.72	53.11	(RDP site)
	5	<i>Caulobacter crescentus</i> CB15	Cau.cres2	67.68	55.08	X52281
	6	<i>Rhodopseudomonas palustris</i>	Rps.palust	64.72	54.92	M59068
	7	<i>Agrobacterium tumefaciens</i> C58 (18)	Ag.tumefac	59.74	54.73	M11223
	8	<i>Brucella melitensis</i> biovar Abortus* (19)	Bru.aborts	57.87	55.40	X13695
	9	<i>Methylobacterium extorquens</i> * (19)	Mlb.extorq	68.80	53.96	M59207
	10	<i>Rhizobium leguminosarum</i> * (18)	Rhb.legumi	60.13	57.02	M63183
Beta	11	<i>Neisseria meningitidis</i> MC58 (12)	Nis.mening	53.06	54.46	Z22776
	12	<i>Ralstonia eutropha</i> (11)	Ral.eutrop	65.53	54.68	M32021
	13	<i>Burkholderia cepacia</i> * (11)	Bur.cepaci	63.77	54.74	M22518
	14	<i>Comamonas testosteroni</i> * (12)	Com.testos	62.89	54.56	M11224
	15	<i>Bordetella pertussis</i> *	Brd.pertus	67.87	55.69	U04950
Gamma	16	<i>Moraxella catarrhalis</i>	Mrx.catarr	42.50	51.55	X74903
	17	<i>Acinetobacter calcoaceticus</i> *	Acn.calcoa	42.34	52.39	M34139
	18	<i>Azotobacter vinelandii</i> * (17)	Azo.vinlnd	65.24	55.38	L40329
	19	<i>Pseudomonas aeruginosa</i> * (17)	Ps.aerugin	65.70	54.18	M34133
	20	<i>Vibrio cholerae</i> *	V.cholerae	47.35	56.65	L05178
	21	<i>Aeromonas hydrophila</i> * (2)	Arm.hydrop	57.32	55.20	M59148
	22	<i>Escherichia coli</i> K12 (13)	E.coli	51.83	54.41	J01695
	23	<i>Buchnera aphidicola</i> * (13)	Buc.aphidi	26.68	49.29	L18927
	24	<i>Salmonella typhimurium</i> LT2 (14)	S.tymurium	53.36	54.34	X80681
	25	<i>Citrobacter freundii</i> * (14)	Cit.freund	50.13	54.56	M59291
	26	<i>Pectobacterium carotovorum</i> * (15)	Er.carotov	50.87	53.70	M59149
	27	<i>Klebsiella pneumoniae</i> (15)	K.pneumoni	55.79	54.63	X87276
	28	<i>Proteus vulgaris</i> * (16)	P.vulgaris	45.43	52.75	X07652
	29	<i>Yersinia pestis</i>	Yer.pestis	48.97	54.93	X67464
Delta/Epsilon	30	<i>Actinobacillus pleuropneumoniae</i> * (16)	Acb.pleuro	39.22	51.91	M75074
	31	<i>Haemophilus influenzae</i> Rd*	H.inflrnrA	38.76	51.98	L42023
	32	<i>Acidithiobacillus ferrooxidans</i>	Thb.ferro2	58.44	56.24	M79430/1/2
	33	<i>Dichelobacter nodosus</i> *	Dch.nodosu	48.41	51.99	M35016
	34	<i>Coxiella burnetii</i>	Cox.burnet	42.52	53.77	M21291
	35	<i>Desulfovibrio vulgaris</i> *	Dsv.vulgar	62.40	56.92	M34399
	36	<i>Myxococcus xanthus</i> *	Myx.xanthu	69.27	55.88	M34114
	37	<i>Helicobacter pylori</i> J99* (10)	Hlb.pylori	39.90	50.42	M88157
	38	<i>Wolinella succinogenes</i> * (10)	Wln.succin	48.58	51.53	M26636
	39	<i>Campylobacter jejuni</i> NCTC 11168* (1)	Cam.jejuni	30.83	49.62	M59298
Other Gram negatives						
	40	<i>Bacteroides fragilis</i>	Bac.fragil	42.07	50.46	M61006
	41	<i>Chlamydomonas pneumoniae</i> AR39*	Clm.pneumo	41.29	49.42	L06108
	42	<i>Plectonema boryanum</i>	Plec.borya	48.28	53.73	(RDP site)
	43	<i>Borrelia burgdorferi</i> *	Bor.burgdo	29.29	45.68	M59293
	44	<i>Leptospira interrogans</i> *	Lps.interR	34.54	52.79	Z12817
Gram positive						
	45	<i>Saccharopolyspora erythraea</i>	Scp.erythr	72.28	58.87	X53198
	46	<i>Amycolatopsis mediterranei</i> *	Amy.medter	72.65	58.77	X76957
	47	<i>Mycobacterium tuberculosis</i> CDC1551 (3)	Myb.tuberc	65.77	57.95	X52917
	48	<i>Corynebacterium glutamicum</i> ATCC 13032 (3)	Cor.glutam	54.72	56.59	Z46753
	49	<i>Rhodococcus erythropolis</i> * (1)	Rco.erythr	63.41	57.60	X53203
	50	<i>Streptomyces coelicolor</i> A3(2)*	Stm.coelic	72.41	58.90	Y00411
	51	<i>Cellulomonas fimi</i> *	Cllm.fimi	72.18	58.17	X83803
	52	<i>Ruminococcus flavefaciens</i> (5)	Ruc.flvfac	49.69	51.15	X85097
	53	<i>Caldicellulosiruptor saccharolyticus</i> (4)	Ccs.saccha	38.02	58.70	L09178
	54	<i>Clostridium botulinum</i> * (4)	C.botulinA	25.13	52.59	X68185
	55	<i>Staphylococcus aureus</i> MW2 (6)	Stp.aureus	33.50	51.00	X68417
	56	<i>Leuconostoc mesenteroides</i> (9)	Lc.mesente	38.52	51.48	M23035
	57	<i>Lactobacillus delbrueckii</i> (9)	L.delbruck	51.48	53.50	M58814
	58	<i>Pediococcus pentosaceus</i> *	Ped.pentos	38.53	50.88	M58834
	59	<i>Lactobacillus casei</i> * (8)	L.casei	44.39	52.72	M23928

TABLE 1. Continued.

	ID	Species	RDP ID	%GC (genomic)	%GC (16s rDNA)	GenBank Accession
	60	<i>Enterococcus faecalis</i> (8)	Eco.faecal	36.78	53.39	(RDP site)
	61	<i>Lactococcus lactis</i> subsp. <i>lactis</i> * (7)	Lcc.lactis	36.10	51.28	M58837
	62	<i>Streptococcus pneumoniae</i> TIGR4* (7)	Stc.pneumo	40.55	53.33	X58312
	63	<i>Bradyrhizobium japonicum</i> (5)	Bb.brevis	62.33	55.00	X60612
	64	<i>Bacillus subtilis</i> * (6)	B.subtilis	44.32	55.09	K00637
	65	<i>Spiroplasma citri</i>	Spp.citri	29.38	49.13	M23942
Proteobacteria (added during revisions)						
Beta	66	<i>Alcaligenes faecalis</i>	Alc.faecal	57.47	54.12	M22508
	67	<i>Rubrivivax gelatinosus</i>	Rub.gelati	69.62	55.76	M60682
	68	<i>Thauera aromatica</i> *	Tha.aromat	62.86	56.47	X77118
Gamma	69	<i>Shigella dysenteriae</i>	Shi.dysntr	46.77	54.56	X80680
	70	<i>Shigella flexneri</i>	Shi.flxner	45.53	54.54	X80679
	71	<i>Shigella sonnei</i>	Shi.sonnei	42.90	54.78	X80726
	72	<i>Salmonella enteritidis</i>	S.enteriti	49.49	54.32	(RDP site)
	73	<i>Erwinia amylovora</i>	Er.amylvor	53.81	56.24	L36466
	74	<i>Pantoea agglomerans</i>	Er.herbico	55.16	54.74	(RDP site)
	75	<i>Serratia marcescens</i>	Ser.marces	57.30	53.87	M59160
	76	<i>Photobacterium luminescens</i>	Pr.lumines	45.38	55.37	X82248
	77	<i>Yersinia enterocolitica</i>	Yer.entero	47.16	53.92	M59292
	78	<i>Yersinia pseudotuberculosis</i>	Yer.ptuber	45.04	54.30	Z21939
		Mean		51.26	54.04	
		Standard deviation		12.43	2.52	

different clades of bacteria. We used ANCML to estimate  $\beta$  from 16S rDNA phylogenies extracted from RDP for five nonoverlapping and unnested groups of bacteria (Fig. 1). These phylogenies are shown in Figures 2a–e, with the y-axis describing GC content as inferred for ancestral nodes by ANCML. All phylogenies show evidence of a strong phylogenetic signal, with closely related species clustering together in GC content, as noted by Ochman and Lawrence (1996). For each phylogeny, we used ANCML to obtain an estimate of  $\beta$ . These estimates, along with their confidence limits based on equation (4), are compared in Figure 4. For the five bacterial clades (Figs. 2a–e), the mean  $\beta$ -value was 997 in units of (%GC)<sup>2</sup> per unit of sequence divergence at the 16S rDNA gene. All confidence intervals are overlapping, indicating that no clade had a significantly different  $\beta$ -value. Thus, there is no evidence that GC content evolves at significantly different rates (i.e., heterogeneous rates) in the bacterial groups illustrated in Figure 1.

To confirm the generality of this result, two additional analyses were performed: one at a broader phylogenetic scale (a mixture of 40 bacterial species drawn randomly from those shown in Figs. 2a–e; see asterisk in Table 1) and one at a more narrow phylogenetic scale (the Enterobacteriaceae group; Fig. 2f).  $\beta$  estimated from the broad mixture of bacteria was 983 (95% CI: 660–1621). Similar values of  $\beta$  were found when this analysis was repeated twice with different random samples of 40 bacteria ( $\beta = 989$  and 1017, respectively).  $\beta$  estimated from the more closely related Enterobacteriaceae group was 1711 (95% CI: 977–3742). The fact that the point estimate for the enterics is higher may be biologically significant, suggesting that GC content evolves more rapidly in this group. Alternatively, the point estimate may be higher simply by chance, which is especially likely given that short branch lengths are estimated with substantial error but have a disproportionately large influence on the

estimation of  $\beta$ . At any rate, all confidence limits overlap (see Fig. 4), suggesting that there is little heterogeneity in the rate of evolution of genomic GC content. These additional analyses are not, however, entirely independent of those described in the previous paragraph, because they include several of the same species (see Table 1).

The confidence intervals for each estimate of  $\beta$  are fairly broad (Fig. 4). For many comparisons of interest, however, the rate at which the standard deviation in GC content rises over time is more relevant than the rate at which the variance increases (see eqs. 2 and 3). The standard deviation in GC content increases over time in proportion to  $\sqrt{\beta}$ . The confidence limits for  $\sqrt{\beta}$  are narrower than those for  $\beta$  simply because taking the square root reduces large numbers; for example, the point estimate for  $\sqrt{\beta}$  is 31.4 with a 95% confidence interval of only 25.7–40.3 for the broad mixture of bacteria. It must be kept in mind, however, that the confidence intervals reported in this paper are based on equation (4) and do not take into account other sources of uncertainty, in particular, uncertainty in the phylogeny, and so they must underestimate the overall confidence intervals.

To determine the time frame over which GC content is likely to remain similar among related bacterial lineages, we inverted equation (2):

$$T_{ai} = 2(\text{SD}_{\text{descendants}})^2/\beta. \quad (6)$$

Given an estimate for  $\beta$  and a criterion for similarity or stationarity in terms of the standard deviation among descendants, equation (6) can be used to specify the time frame over which bacteria will remain similar. For similarity, we use the criterion that 95% of descendants of an ancestor are within  $\zeta$  %GC of the ancestral state. The appropriate value of  $\zeta$  depends on the process or application of interest. If the process is very sensitive to changes in GC content, then a stringent criterion should be chosen (low  $\zeta$ ). Conversely, if

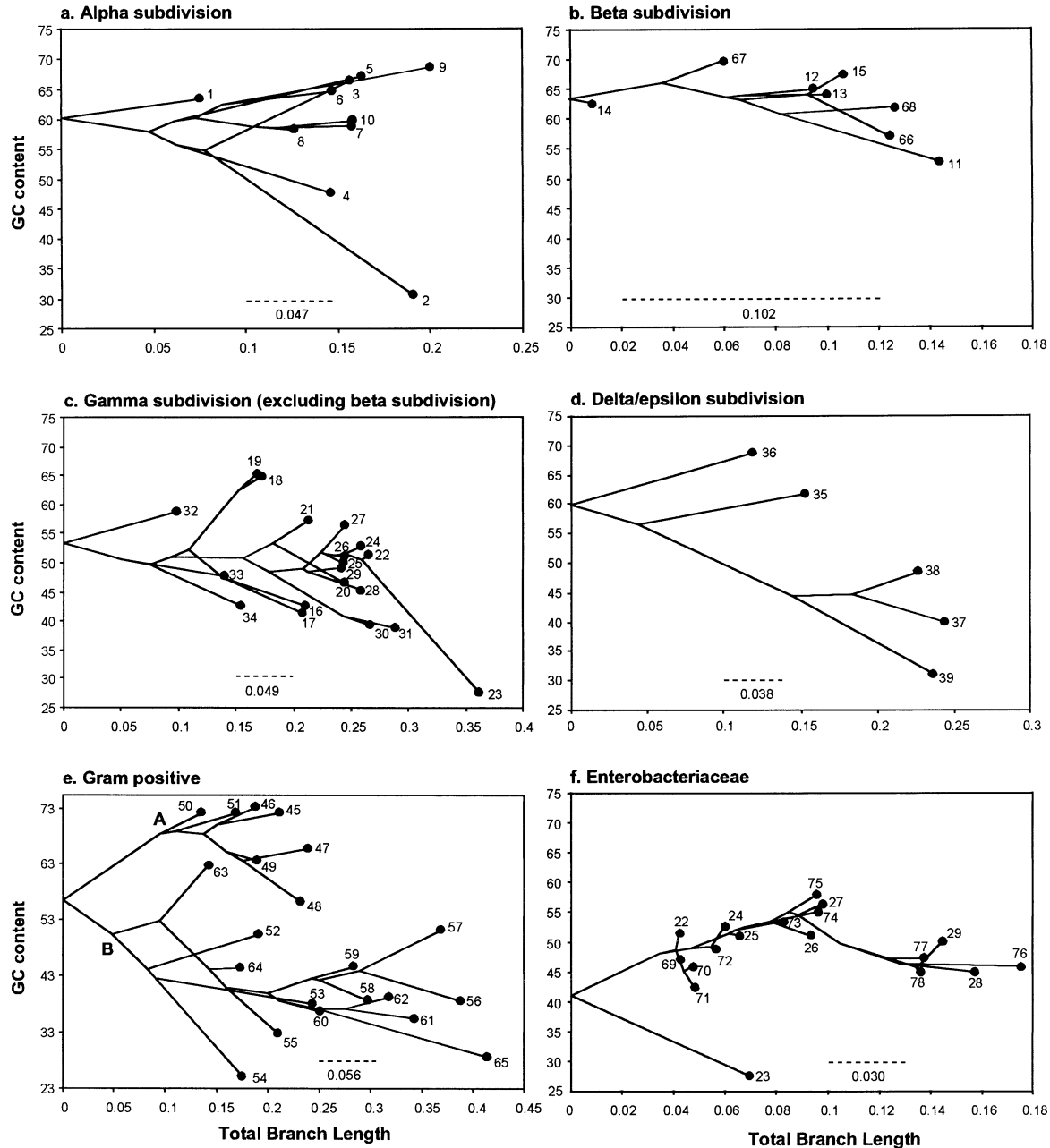


FIG. 2. Phylogeny plots illustrating GC content evolution. The x-axis indicates divergence times between descendants and ancestors based on the 16S rDNA phylogeny. The y-axis indicates GC content, either observed among extant taxa (identified by number in Table 1) or inferred for ancestral nodes using ANCML (SE available upon request). Figures a–d illustrate GC content evolution among subdivisions of the proteobacteria: (a) alpha; (b) beta; (c) gamma; (d) delta/epsilon; (e) Gram positives; (f) Enterobacteriaceae group of gamma bacteria. The dashed lines represent the time scale over which 95% of descendants fall within  $\pm 10\%$  GC of the ancestor using the  $\beta$ -value obtained from each group (see Fig. 4).

only large changes in GC content are likely to affect the process, then a relaxed criterion may be employed (large  $\zeta$ ). Using  $\beta$  estimated from the broad mixture of bacteria ( $\beta = 983$ ), the time period over which descendants remain similar in GC content to their ancestor equals  $0.00053 \zeta^2$  (95% CI based on the CI for  $\beta$ :  $0.00032\zeta^2$ – $0.00079\zeta^2$ ) in units of sequence divergence (substitutions/site at the 16S rDNA gene). Thus, the average period during which 95% of descendants have a %GC value within  $\pm 1\%$  GC of their ancestral value

is very short: 0.00053 substitutions/site. On the other hand, 95% of descendants will remain within 10 %GC of their ancestral value for substantially longer: 0.053 substitutions/site. For scale, the inferred number of substitutions per site in the 16S rDNA gene was, on average, 0.017 since the most recent common ancestor for *Escherichia coli* and *Salmonella typhimurium*. Over this time frame, we expect that 95% of descendants will remain within 6 %GC of their ancestral value.

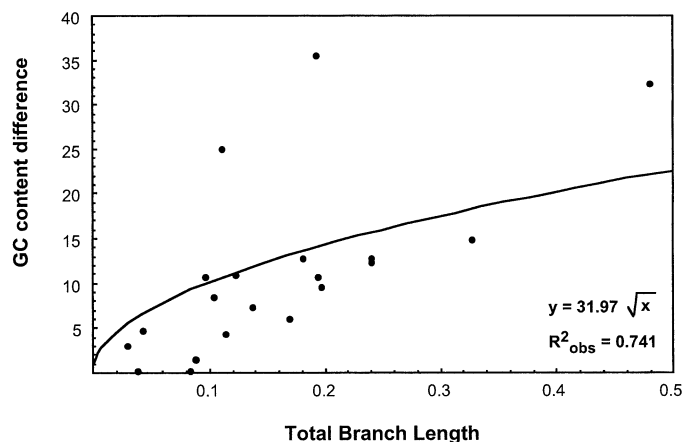


FIG. 3. The relationship between the absolute difference in GC content between two species and the total branch lengths separating them on the 16S rDNA phylogeny. Twenty phylogenetically independent pairs of species are compared (listed in Table 1). The best-fitting square-root function to the data ( $Y_i = b[X_i]^{1/2}$ ) is shown as a solid curve.

A potential source of error in these analyses is that the sequenced protein-coding genes used to determine genomic GC content could be unrepresentative of their host genomes. To check this effect, we calculated  $\beta$  using total genomic GC content from 19 bacteria with completely sequenced genomes. The estimate of  $\beta$  was 955 based on the GC content of protein-coding genes (Nakamura et al. 1997; <http://www.kazusa.or.jp/codon/>) and 903 based on total genomic GC content ([http://www.genome.ad.jp/kegg/catalog/org\\_list.html](http://www.genome.ad.jp/kegg/catalog/org_list.html)). We thus conclude that using data from incompletely sequenced genomes has not substantially biased our results.

Another source of error is that we have used a Brownian-motion model that ignores constraints on GC content. Genomic GC content is certainly constrained to fall between 0% and 100% and appears to be constrained to remain be-

tween 20% and 80% (Table 1). Selection on GC content must increase as these extreme levels are approached because the spectrum of amino acids that can be coded within a genome becomes more limited. To test the effect of these constraints, the data on the broad mixture of bacteria (see asterisk in Table 1) were reanalyzed using an arcsine-root transformation. The arcsine-root transformation is appropriate for data that lie between an upper and lower bound (Zar 1996). In this case, for data constrained within a range of 20% to 80% GC, we used the transformation:

$$GC_{\text{transformed}} = \arcsin \left[ \left( \frac{GC_{\text{raw}} - 20}{80 - 20} \right)^{1/2} \right]. \quad (7)$$

Equation (7) has the effect of increasing the apparent differences between species whose GC contents approach 20% or 80%. Nevertheless, the  $\beta$  value obtained through an ANCMML analysis of the transformed data predicted similar rates of GC content evolution for species with intermediate GC content. For example, consider an ancestral species whose GC content is 50%; using  $\beta$  estimated from the transformed data, we would expect 95% of its descendants to have GC content values within 35.6% and 64.4% after an evolutionary period corresponding to 0.1 substitutions per site in the 16S rDNA gene. This range is very similar to that obtained from the untransformed data analysis (36.3–63.7%). Thus, adding boundary constraints had little effect on the estimated rate of evolution away from the boundaries.

#### DISCUSSION

Many evolutionary mechanisms have been proposed to explain GC content diversity among bacteria, but most authors agree that a species' genomic GC content is set by a balance between selective constraints at the level of codons and amino acids and directional mutational pressure at the nucleotide level (Sueoka 1993; Galtier and Lobry 1997). In this paper, we have used a Brownian-motion model to approximate the

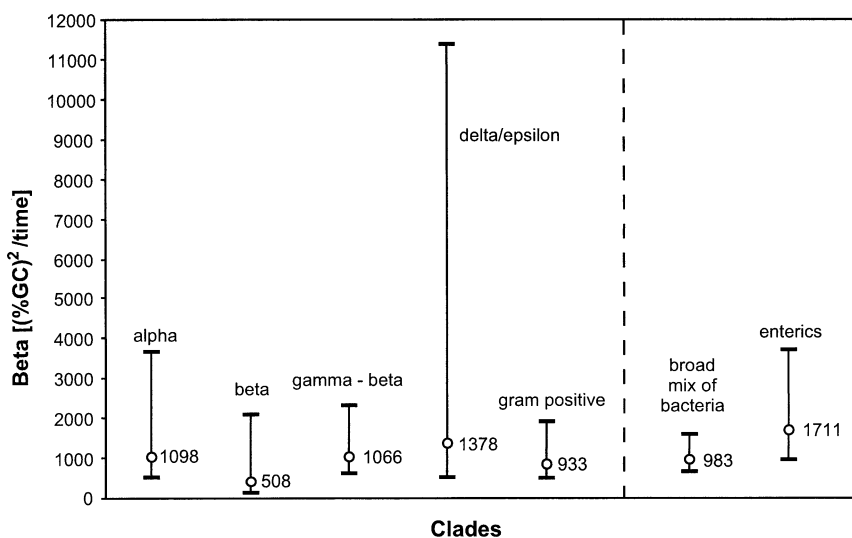


FIG. 4.  $\beta$ -values, with confidence intervals, calculated for the phylogenies shown in Figure 2. The  $\beta$ -value averaged over the five independent phylogenies (Figs. 2a–e) is 997 (change in %GC)<sup>2</sup>/nucleotide substitution. The species included in the broad mixture of bacteria are indicated by an asterisk in Table 1. The enteric species are as described in Figure 2f.

evolution of GC content over time. This model assumes that GC content changes over time according to a time-homogeneous diffusion process, such that the expected GC content of descendant taxa equals the GC content of their ancestors but with variation around this expectation. The use of a Brownian-motion model does not assume that the character is neutral, rather the model is used to approximate both neutral and selective changes in a character, as long as the above assumptions hold approximately (Schluter et al. 1997).

An analysis of changes in GC content as a function of divergence time revealed features consistent with a Brownian-motion model. A Brownian-motion model predicts that the absolute difference in GC content should rise as the square root of branch length,  $Y_i = b(X_i)^{1/2}$ . Such a relationship fits data from 20 species pairs well (Fig. 3); the fit was not significantly improved by the addition to the model of a nonzero intercept or by a linear term ( $X_i$ ). In addition, the amount of scatter around this relationship ( $R_{\text{obs}}^2$ ) was consistent with the amount expected under a Brownian-motion model based on a parametric bootstrap test. Furthermore, we found that GC content has evolved at an average rate of  $\beta \approx 1000$  (change in %GC)<sup>2</sup>/(nucleotide substitution at the 16S rDNA gene) in five independent clades of bacteria (Table 1; Figs. 2a–e). None of the rates estimated differed significantly from any other (Fig. 4), indicating that the rate of GC content change is remarkably homogeneous among diverse groups of bacteria. Thus, using three different criteria, we have found that a Brownian-motion model provides a good description of GC content evolution over the time scales investigated.

Of course, failing to reject a model does not mean that the model is true. Indeed, it must be the case that, at some level, genomic GC content changes in a heterogeneous fashion. Horizontal transfer can cause rather sudden shifts in GC content; mutation rates can change abruptly due to alterations of the repair machinery of a cell; and different bacterial lineages are almost certainly subject to persistent differences in selection. Nevertheless, over long evolutionary time scales, these sources of heterogeneity are not sufficiently pronounced to warrant rejection of the homogeneous Brownian-motion model. If we better understood how the rate of GC content evolution varies over time, we could construct a more complex time-heterogeneous diffusion model for  $\beta$ . Currently, however, this would require us to make ad hoc assumptions about how the rate of %GC change might vary over time and among species. Given that we fail to reject a Brownian-motion model with a homogeneous rate of GC content change, we recommend that it be used as a working model for describing GC content evolution. This model fits the data far better than the alternative hypothesis that the forces shaping GC content (i.e.,  $\gamma$ ) have remained unchanged over time, as has often been assumed in the past (see below).

Using the Brownian-motion model as a null model can help us to identify when unusual changes in GC content have occurred. To conclude that GC content has undergone an unusual evolutionary shift in a group of bacteria, one should first reject the hypothesis that GC content change has been evolving in the group at a normal rate under the Brownian-motion model. For example, consider the well-known division of Gram-positive bacteria into high- and low-GC groups (clades to the right of nodes A and B in Fig. 2e). This basal

division in the Gram positives represents a total divergence of 0.14 substitutions/site and a change in GC content of 18.3% (SE = 6.8). One straightforward explanation for this divergence is that the ancestors of the two groups underwent the same process of GC content evolution seen throughout the bacterial phylogeny. Using equation (2), the overall average  $\beta$ -value (1000), and properties of a normal distribution, we would expect such a large or larger divergence in GC content 3.0% of the time. Although this represents a probability less than 0.05, there are two reasons to suspect that the divergence is not significantly different than expected. First, there is a multiple-comparisons problem, in that many clades of bacteria have been studied, and this pair of clades was considered to be particularly unusual. Second, a marginally higher  $\beta$ -value (>1230) would predict such a high GC content divergence more than 5% of the time. We thus conclude that there is not sufficient evidence to invoke a period of unusually rapid GC content evolution early in the evolution of Gram-positive bacteria.

As a second example, consider the two obligately intracellular bacteria, *Rickettsia* (2 Fig. 2a) and *Buchnera* (23 Fig. 2c). Using ANCMML, there is an inferred %GC content change of 24.2 (SE = 5.0) over a branch of length 0.113 leading to *Rickettsia* and a %GC content change of 23.0 (SE = 2.3) over a branch of length 0.103 leading to *Buchnera*. In both cases, the probability of seeing such a dramatic change based on the overall average  $\beta$ -value (1000) is less than 0.002. These results remain significant following a Bonferroni correction for multiple comparisons as long as comparisons are restricted to the obligately intracellular bacteria (*Buchnera aphidicola*, *Rickettsia prowazekii*, *Coxiella burnetti*, and *Chlamydomydia pneumoniae*) and not to all 78 bacteria included in this study. Furthermore, a substantially higher  $\beta$ -value (> 6900) is needed to see such dramatic changes in two independent comparisons at least 5% of the time. Thus, there is good evidence that the evolutionary history of obligately intracellular bacteria has involved unusually rapid rates of GC content evolution. A general trend toward low GC content in intracellular bacteria has been noted previously (Moran 1996; Heddi et al. 1998). Moran (1996) concluded that mutational bias combined with Muller's ratchet within the small, asexual populations of intracellular bacteria was most likely responsible for the rapid evolution of GC content in *Buchnera* spp., based on a study of synonymous and nonsynonymous substitutions. (As an aside, if the gamma [or alpha] subgroup of eubacteria had exhibited, as a whole, the elevated rate of GC content evolution observed in *Buchnera* [or *Rickettsia*], the null hypothesis that the rates of GC content change are the same in the different groups of bacteria would have been rejected; see Fig. 4.)

The Brownian-motion model also allows us to evaluate the changes that are likely to occur in GC content over a particular time frame. We have found that, although the GC content of related taxa remains correlated for long periods of time, GC content may be considered stationary only over short periods. In particular, one may expect GC content to remain within  $\pm 1$  %GC only among organisms whose 16S rDNA sequences differ from their ancestor by less than about one in 2000 nucleotides (using the overall average  $\beta$ -value of 1000). On the other hand, organisms that are much more

distantly related to their ancestor, with as many as one in 20 substitutions per 16S rDNA site, still have a 95% chance of being within  $\pm 10\%$  GC of each other. Following Lawrence and Ochman (1997), we can use the estimated time of 100 million years since the divergence of *Escherichia coli* and *Salmonella* spp. from their common ancestor to convert branch lengths into time estimates. Using this conversion (i.e., 0.017 substitutions per nucleotide per 100 million years), 95% of the descendants are expected to have percent GC values within  $\pm 1$  of the ancestral state over 3 million years and within  $\pm 10\%$  over 300 million years. This conclusion agrees qualitatively with that of Rzhetsky and Nei (1995), who used a linear invariant method to test statistically the assumption of stationarity in base composition and found that the assumption of stationarity was rejected for all but very closely related sequences.

Knowing the extent to which GC content is expected to change over a given time period can help determine whether these changes should be taken into account in a study. Currently, many analyses of sequence data ignore underlying changes in genomic GC content. For example, an influential model of horizontal gene transfer assumes that GC content is a static property of bacterial lineages and tracks how GC content in a transferred sequence evolves to match that of its new host genome (Lawrence and Ochman 1997, 1998). Similarly, most models of evolutionary change used in phylogenetic reconstruction (e.g., Jukes-Cantor, Kimura two-parameter, Tamura, Tajima-Nei) assume that GC content is either stationary and/or uniform across taxa (Swofford et al. 1996; Galtier and Gouy 1998). Nonhomogeneous base composition presents a problem for phylogenetic reconstruction (Galtier and Gouy 1998), which can be lessened by selecting taxa of similar base composition. This technique does not, however, ensure that the ancestors of these taxa also had the same GC content. As is evident in Figure 2, the eventual randomization with respect to phylogeny means that two species of similar base composition can have very different GC content histories (e.g., taxa 3 and 5 in Fig. 2a). This could bias any model of evolution that assumed stationarity. To assess this bias, one could determine the amount of GC content change expected under the Brownian-motion model and assess the sensitivity of an analysis to such changes in GC content. For example, the Brownian-motion model can be used to generate sequence data that can be used to test the sensitivity of phylogenetic models that ignore changing GC content. For deep phylogenetic trees, %GC is likely to change extensively, and methods that take such changes into account are more likely to provide accurate phylogenetic reconstructions (e.g., Lockhart et al. 1994; Galtier and Gouy 1998).

In conclusion, we have shown that GC content evolves over time in a manner that is consistent with a time-homogeneous model of Brownian motion, with variance in %GC content rising at an approximate rate of  $1000 (\%GC)^2 / (\text{nucleotide substitution at the 16S rDNA gene})$ . This indicates that the rate of GC content evolution is neither so slow nor so fast that it may be ignored, except perhaps in studies of very closely related bacteria. Similar studies, using measures other than GC content, using different assemblages of species, and/or incorporating uncertainty in the phylogeny,

promise to refine our understanding of the tempo and mode by which genomic base composition evolves.

#### ACKNOWLEDGMENTS

We are grateful to F. Brinkman, N. Galtier, D. Schluter, M. Whitlock, and several anonymous reviewers for helpful comments and suggestions and to J. Sunday for help with the data collection. This work was funded by grants to SPO from the Natural Sciences and Engineering Research Council (Canada), the Peter Wall Institute for Advanced Studies, and the Centre National de la Recherche Scientifique (France).

#### LITERATURE CITED

- Chiusano, M. L., G. D'Onofrio, F. Alvarez-Valin, K. Jabbari, G. Colonna, and G. Bernardi. 1999. Correlations of nucleotide substitution rates and base composition of mammalian coding sequences with protein structure. *Gene* 238:23–31.
- D'Onofrio, G., K. Jabbari, H. Musto, and G. Bernardi. 1999. The correlation of protein hydrophobicity with the base composition of coding sequences. *Gene* 238:3–14.
- Drake, J. W. 1991. A constant rate of spontaneous mutation in DNA-based microbes. *Proc. Natl. Acad. Sci. USA* 88:7160–7164.
- Felsenstein, J. 1981. Evolutionary trees from DNA sequences: a maximum likelihood approach. *J. Mol. Evol.* 17:368–376.
- . 1985. Phylogenies and the comparative method. *Am. Nat.* 125:1–15.
- . 1993. PHYLIP (phylogeny inference package). Ver. 3.5c. Dept. of Genetics, Univ. of Washington, Seattle, WA.
- Galtier, N., and M. Gouy. 1995. Inferring phylogenies from sequences of unequal base compositions. *Proc. Natl. Acad. Sci. USA* 92:11317–11321.
- . 1998. Inferring pattern and process: maximum-likelihood implementation of a nonhomogeneous model of DNA sequence evolution for phylogenetic analysis. *Mol. Biol. Evol.* 15: 871–879.
- Galtier, N., and J. R. Lobry. 1997. Relationships between genomic G+C content, RNA secondary structures and optimal growth temperature in prokaryotes. *J. Mol. Evol.* 44:632–636.
- Gu, X., D. Hewett-Emmett, and W.-H. Li. 1998. Directional mutational pressure affects the amino acid composition and hydrophobicity of proteins in bacteria. *Genetica* 102/103:383–391.
- Heddi, A., H. Charles, C. Khatchadourian, G. Bonnot, and P. Nardon. 1998. Molecular characterization of the principle symbiotic bacteria of the weevil *Sitophilus oryzae*: a peculiar G+C content of an endocytobiotic DNA. *J. Mol. Evol.* 47:52–61.
- Lawrence, J. G., and H. Ochman. 1997. Amelioration of bacterial genomes: rates of change and exchange. *J. Mol. Evol.* 44: 383–397.
- . 1998. Molecular archaeology of the *Escherichia coli* genome. *Proc. Natl. Acad. Sci. USA* 95:9413–9417.
- 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.
- Ludwig, W., O. Strunk, S. Klugbauer, N. Klugbauer, M. Weizenegger, J. Neumaier, M. Bachleitner, and K. H. Schleifer. 1998. Bacterial phylogeny based on comparative sequence analysis. *Electrophoresis* 19:554–568.
- Maidak, B. L., G. J. Olsen, N. Larsen, R. Overbeek, M. J. McCaughey, and C. R. Woese. 1997. The RDP (Ribosomal Database Project). *Nucleic Acids Res.* 25:109–111.
- Moran, N. A. 1996. Accelerated evolution and Muller's ratchet in endosymbiotic bacteria. *Proc. Natl. Acad. Sci. USA* 93: 2873–2878.
- Nakamura, Y., T. Gojobori, and T. Ikemura. 1997. Codon usage tabulated from the international DNA sequence databases. *Nucleic Acids Res.* 25:244–245.
- Ochman, H., and J. G. Lawrence. 1996. Phylogenetics and the amelioration of bacterial genomes. Pp. 2627–2637 in F. C. Neidhardt, R. Curtiss III, J. L. Ingraham, E. C. C. Lin, K. B. Low, B.

- Magasanik, W. S. Reznikoff, M. Riley, M. Schaechter, and H. E. Umbarger, eds. *Escherichia coli* and *Salmonella*: cellular and molecular biology. 2d ed. American Society for Microbiology, Washington, D.C.
- Olsen, G. J., H. Matsuda, R. Hagstrom, and R. Overbeek. 1994. fastDNAm1: a tool for construction of phylogenetic trees of DNA sequences using maximum likelihood. *Comput. Appl. Biosci.* 10:41–48.
- Pan, A., C. Dutta, and J. Das. 1998. Codon usage in highly expressed genes of *Haemophilus influenzae* and *Mycobacterium tuberculosis*: translational selection versus mutational bias. *Gene* 215: 405–413.
- Rzhetsky, A., and M. Nei. 1995. Tests of applicability of several substitution models for DNA sequence data. *Mol. Biol. Evol.* 12:131–151.
- Schluter, D., T. Price, A. Ø. Mooers, and D. Ludwig. 1997. Likelihood of ancestor states in adaptive radiation. *Evolution* 51: 1699–1711.
- Shields, D. C. 1990. Switches in species-specific codon preferences: the influence of mutation biases. *J. Mol. Evol.* 31:71–80.
- Sueoka, N. 1993. Directional mutation pressure, mutator mutations and dynamics of molecular evolution. *J. Mol. Evol.* 37:137–153.
- Swofford, D. L., G. J. Olsen, P. J. Waddell, and D. M. Hillis. 1996. Phylogenetic Inference. Pp. 407–514 in D. M. Hillis, C. Moritz, and B. K. Mable, eds. *Molecular systematics*. 2d ed. Sinauer Associates, Sunderland, MA.
- Taylor, H. M., and S. Karlin. 1998. An introduction to stochastic modeling. 3rd ed. Academic Press, San Diego, CA.
- Wolfram, S. 1991. *Mathematica*. Addison Wesley, New York.
- Zar, J. H. 1996. *Biostatistical analysis*. 3rd ed. Prentice-Hall, Englewood Cliffs, NJ.

Corresponding Editor: J. Huelsenbeck