Copyright © 2004 by the Genetics Society of America
DOI: 10.1534/genetics.104.026344

# Genome-Wide Patterns of Nucleotide Substitution Reveal Stringent Functional Constraints on the Protein Sequences of Thermophiles

**Robert Friedman,* John W. Drake† and Austin L. Hughes*,[1]**

**Department of Biological Sciences, University of South Carolina, Columbia, South Carolina 29208 and †Laboratory of Molecular Genetics, National Institute of Environmental Health Sciences, Research Triangle Park, North Carolina 27709-2233*

Manuscript received January 8, 2004
Accepted for publication April 16, 2004

ABSTRACT

To test the hypothesis that the proteins of thermophilic prokaryotes are subject to unusually stringent functional constraints, we estimated the numbers of synonymous and nonsynonymous nucleotide substitutions per site between 17,957 pairs of orthologous genes from 22 pairs of closely related species of Archaea and Bacteria. The average ratio of nonsynonymous to synonymous substitutions was significantly lower in thermophiles than in nonthermophiles, and this effect was observed in both Archaea and Bacteria. There was no evidence that this difference could be explained by factors such as nucleotide content bias. Rather, the results support the hypothesis that proteins of thermophiles are subject to unusually strong purifying selection, leading to a reduced overall level of amino acid evolution per mutational event. The results show that genome-wide patterns of sequence evolution can be influenced by natural selection exerted by a species' environment and shed light on a previous observation that relatively few of the mutations arising in a thermophilic archaeon were nucleotide substitutions in contrast to indels.

HIGHLY thermophilic species, inhabiting environments where temperatures can exceed 100°, have been described for both domains of prokaryotic life (Woese *et al.* 1990), the Bacteria (or eubacteria) and the Archaea (or archaebacteria). Considerable interest has focused on understanding the mechanisms that make life possible under these extreme conditions (Brown and Lupas 1998). Both analyses of protein structures and genome-wide comparisons of amino acid composition have revealed widely shared characteristics of proteins encoded by the genomes of thermophiles; these include the presence of surface ion networks, location of hydrophobic residues in partly surface-exposed positions, avoidance of thermally unstable amino acid residues, and overall compact protein structure (Perutz 1978; Brown and Lupas 1998; Gromiha *et al.* 1999; Cambillau and Claverie 2000; Kumar *et al.* 2000; Kumar and Nussinov 2001; Chen *et al.* 2003; Criswell *et al.* 2003). In addition, there is evidence that the proteins of thermophiles are characterized by a distinct pattern of amino acid composition (Kreil and Ouzounis 2001; Tekaia *et al.* 2002). The evident survival value conferred by adaptive protein features is expected to result in strong purifying (conservative) natural selection on protein-coding genes of thermophiles.

We tested this prediction by estimating the numbers of synonymous and nonsynonymous (amino acid-altering) nucleotide substitutions by comparing 17,957 pairs of orthologous gene pairs from closely related species of Bacteria and Archaea. We compared 16 pairs of closely related Eubacteria (usually congeners) whose genomes have been completely sequenced, completely sequenced genomes from three genera of Archaea (including the thermophiles Pyrococcus and Sulfolobus), and sequences from partially sequenced genomes of three thermophilic genera of Bacteria (Thermotoga, Thermoanaerobacter, and Thermus).

We estimated the number of synonymous nucleotide substitutions per synonymous site ($d_S$) and the number of nonsynonymous nucleotide substitutions per nonsynonymous site ($d_N$) between orthologous gene pairs. Because synonymous mutations are less likely than nonsynonymous mutations to be selectively deleterious, $d_S$ between two related sequences is expected to reflect both the mutation rate and the time since the sequences' last common ancestor (Nei 1987). By contrast, $d_N$ will reflect mainly the effects of purifying selection, which acts to eliminate deleterious mutations. Thus, the $d_N/d_S$ ratio is a measure of the strength of purifying selection at the amino acid level, a lower ratio implying stronger selective constraint.

## METHODS

**Sequences analyzed:** We compared orthologous gene pairs identified in 16 complete genomes of Bacteria and three complete genomes of Archaea (Table 1). In addition, available sequences were compared within three genera of thermophilic Bacteria for which com-

[1]*Corresponding author:* Department of Biological Sciences, University of South Carolina, 700 Sumter St., Columbia, SC 29208. E-mail: austin@biol.sc.edu

**TABLE 1**

**Sequences used in analyses**

| Domain | Group | Comparison | No. of orthologous gene pairs | $d_N/d_S$[a] | GC3 (%)[a] |
|---|---|---|---|---|---|
| Bacteria | Mesophiles | *Bacillus halodurans* C-125 *vs. B. subtilis* | 968 | 0.541 ± 0.011 | 52.3 ± 0.2 |
| | | *Brucella melitensis vs. B. suis* | 472 | 0.408 ± 0.021 | 66.6 ± 0.3 |
| | | *Buchnera aphidicola vs. Buchnera* sp. APs | 354 | 0.101 ± 0.004 | 14.9 ± 0.1 |
| | | *Clostridium acetobulyticum vs. C. perfringens* | 593 | 0.121 ± 0.003 | 18.8 ± 0.1 |
| | | *Corynebacterium efficiens* YS-314 *vs. C. glutamicum* | 1259 | 0.088 ± 0.002 | 70.4 ± 0.1 |
| | | *Escherichia coli* K12 *vs. Salmonella typhimurium* LT2 | 1537 | 0.088 ± 0.003 | 57.1 ± 0.2 |
| | | *Listeria monocytogenes* EGD-e *vs. L. innocua* | 1355 | 0.054 ± 0.002 | 29.6 ± 0.1 |
| | | *Mycobacterium leprae vs. M. tuberculosis* H37Rv | 809 | 0.103 ± 0.003 | 74.8 ± 0.2 |
| | | *Mycoplasma genitalium vs. M. pneumoniae* | 103 | 0.085 ± 0.005 | 33.5 ± 0.6 |
| | | *Neisseria meningitidis* MC58 *vs. N. meningitidis* Z2491 | 1041 | 0.186 ± 0.009 | 63.2 ± 0.3 |
| | | *Pseudomonas aeruginosa vs. P. putida* KT2440 | 1357 | 0.112 ± 0.002 | 82.4 ± 0.2 |
| | | *Rickettsia conorii* Malish 7 *vs. R. prozawekii* | 612 | 0.125 ± 0.004 | 20.6 ± 0.1 |
| | | *Staphylococcus aureus* MW2 *vs. S. aureus* N315 | 939 | 0.138 ± 0.008 | 22.8 ± 0.1 |
| | | *Streptococcus agalactiae* 2603V/R *vs. S. pyogenes* | 729 | 0.092 ± 0.003 | 28.2 ± 0.2 |
| | | *Tropheryma whipplei* str. Twist *vs. T. whipplei* str. TW08/27 | 308 | 0.216 ± 0.002 | 42.7 ± 0.3 |
| | | *Xanthomonas axonopodis* p. citri str. 306 *vs. X. campestris* | 1845 | 0.077 ± 0.002 | 81.8 ± 0.1 |
| | Thermophiles | *Thermoanaerobacter ethanolicus vs. T. tengcongensis* | 9 | 0.124 ± 0.023 | 32.3 ± 1.0 |
| | | *Thermotoga maritima vs. T. neapolitana* | 64 | 0.091 ± 0.009 | 54.7 ± 0.6 |
| | | *Thermus aquaticus vs. T. thermus* | 16 | 0.073 ± 0.019 | 91.4 ± 1.7 |
| Archaea | Mesophiles | *Methanosarcina acetivorans* str. C2A *vs. M. mazei* Goe | 1452 | 0.131 ± 0.002 | 47.5 ± 0.2 |
| | Thermophiles | *Pyrococcus abyssi vs. P. horokoshii* | 1126 | 0.086 ± 0.002 | 46.9 ± 0.2 |
| | | *Sulfolobus solfataricus vs. S. tokodaii* | 1009 | 0.085 ± 0.001 | 25.9 ± 0.1 |

[a] Mean ± standard error.

plete genomes were not available (Table 1). We identified orthologous gene pairs (genes homologous by descent from a common ancestral gene without gene duplication) between each of these pairs of species. We applied the BLASTCLUST computer program, available from the collection of BLAST tools (Altschul *et al.* 1997), to cluster protein translations to define protein families (single-linkage method). Homology search between sequences was performed using an *E* value of $10^{-50}$ and a minimum of 20% identity across at least 30% of the sequence lengths. These strict search criteria were used to increase our chance of identifying orthologous gene pairs rather than paralogues. Only families with one member from each species in the pair were used in analyses, to avoid the problem of paralogous comparisons. Sequences were aligned at the amino acid level using the CLUSTALW program (Thompson *et al.* 1994) and the alignment was imposed on the coding sequences.

**Evolutionary analyses:** The maximum-likelihood (ML) method (Yang and Nielsen 2000) implemented in the PAML program (Yang 1997) was used to estimate the number of synonymous nucleotide substitutions per synonymous site ($d_S$) and the number of nonsynonymous nucleotide differences per nonsynonymous site ($d_N$). We used the F3X4 model, which incorporates the nucleotide composition and transition/transversion ratios estimated from the compared sequences and assumes equal likelihood of each possible codon pathway. We also estimated $d_S$ and $d_N$ by the modified Nei-Gojobori method (Zhang *et al.* 1998); because the results of both methods were similar, only the ML results are presented here. We excluded from analyses those cases in which there were no synonymous differences or where $d_S$ was undefined; of 19,606 comparisons between putative orthologs, 1649 (8.4%) were excluded on these grounds.

We estimated $d_S$ and $d_N$ between putative orthologs in pairs of closely related species (Table 1). Note that because these comparisons were between orthologous genes from pairs of closely related species, each nucleotide difference between pair members must have arisen since the most recent common ancestor of the two species. Thus, each comparison is phylogenetically and statistically independent of each other comparison (Felsenstein 1985). In statistical analyses, we assumed that each orthologous gene pair evolves independently of other genes in the genome. However, because we used the orthologous gene pair as the unit of statistical analysis, our analyses did not require the assumption that each nucleotide site evolves independently, a biologically unrealistic assumption that is routinely made in studies of molecular evolution.

We also used phylogenetically independent comparisons in the analysis of nucleotide content. Nucleotide content (both G + C content and A + G content) was measured at third codon positions in each of the genes compared, and then nucleotide content was averaged

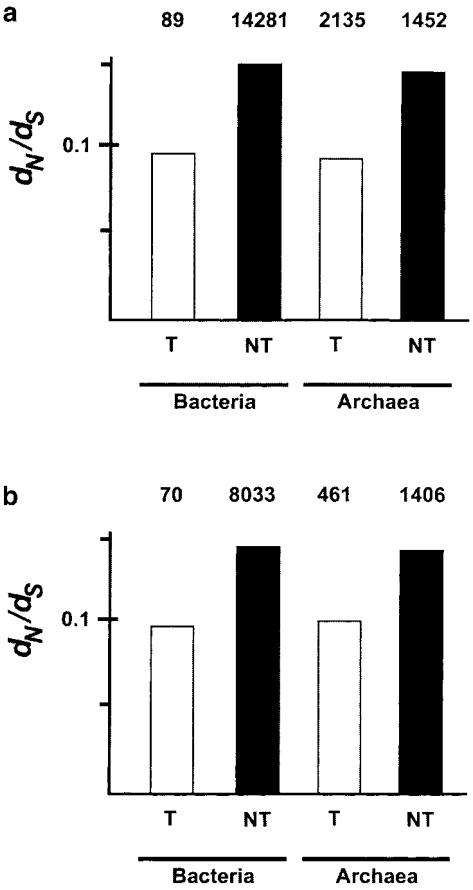


FIGURE 1.—(a) Mean $d_N/d_S$ ratio in comparisons of orthologous genes between closely related species pairs of thermophilic (T) and nonthermophilic (NT) Bacteria and Archaea. Numbers of comparisons in each category are shown. (b) Mean $d_N/d_S$ in comparisons of orthologous genes excluding cases with $d_S$ or $d_N \geq 1.5$.

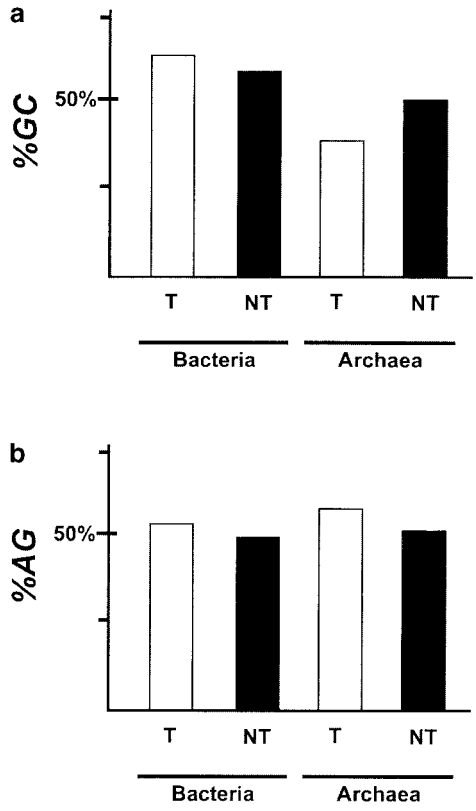


FIGURE 2.—(a) Mean percentage of G + C at third codon positions (GC3) in genes used in analyses. (b) Mean percentage of A + G at third codon positions (AG3) in genes used in analyses. T, thermophilic species; NT, nonthermophilic species.

between the two members of each orthologous pair. Thus values of G + C at the third position (GC3) and of A + G at the third position (AG3) were obtained for each orthologous gene pair.

## RESULTS

Factorial analysis of variance in mean $d_N/d_S$ ratio showed no significant difference between the two prokaryotic domains (Bacteria and Archaea) but a significant difference between thermophiles and nonthermophiles ($F_{1,17956} = 19.09$; $P < 0.001$; Figure 1a). There was no significant interaction between domain and thermophily, indicating a similar difference between thermophiles and nonthermophiles in the two prokaryotic domains.

Although the ML method (YANG and NIELSEN 2000) corrects for nucleotide content bias, it is well known that under simple models, nucleotide content bias at third-codon positions can affect the observed numbers of synonymous substitutions (WOLFE *et al.* 1989). If nucleotide-content bias had a similar effect in our analyses, the higher $d_N/d_S$ ratio we observed in nonthermophiles might be simply due to a reduction in observed $d_S$ resulting from compositional bias in nonthermophile species. To test whether such a phenomenon could be responsible for our results, we examined the percentage of G + C at third-codon positions (GC3) in the genes in our data set (Figure 2a). Factorial analysis of variance showed a significant difference in mean GC3 between the two prokaryotic domains ($F_{1,17956} = 142.27$; $P < 0.001$) and a significant difference between thermophiles and nonthermophiles ($F_{1,17956} = 5.76$; $P = 0.016$). There was also a significant interaction between domain and thermophily ($F_{1,17956} = 40.28$; $P < 0.001$).

These results are explained by the observation that mean GC3 was higher in thermophiles than in nonthermophiles in the case of Bacteria, but lower in thermophiles than in nonthermophiles in the case of Archaea (Figure 2a). In both Bacteria and Archaea, GC3 was closer on average to 50% in nonthermophiles than in thermophiles (Figure 2a). In the entire data set, there was a modest but significant negative correlation between $d_S$ and GC3 ($r = -0.101$; $P < 0.001$). Thus, on the basis of G + C content alone, the $d_N/d_S$ ratio would be expected to be highest in thermophilic Bacteria, intermediate in nonthermophiles of both domains, and

lowest in thermophilic Archaea. The fact that this pattern was not observed (Figure 1) supports the hypothesis that G + C content bias was not a major factor in yielding the observed pattern in $d_N/d_S$ ratios. When GC3 was included as a covariate in the analysis of variance in $d_N/d_S$, there was a significant effect of the covariate ($F_{1,17956}$ = 5.10; $P = 0.024$), but the significant effect of thermophily remained ($F_{1,17956}$ = 19.44; $P < 0.001$). This result indicates that the association between thermophily and a reduced $d_N/d_S$ was statistically independent of the linear relationship between $d_N/d_S$ and GC3.

In the case of thermophiles, some data suggest a preference for purines in mRNAs (Lao and Forsdyke 2000; Lambros *et al.* 2003). On the hypothesis that a bias in nucleotide content is most likely to be expressed at third-codon positions, we examined the percentage of A + G at third-codon positions (AG3) in the genes in our data set (Figure 2b). Factorial analysis of variance showed a significant difference in mean AG3 between the two prokaryotic domains ($F_{1,17956} = 60.53$; $P < 0.001$) and a significant difference between thermophiles and nonthermophiles ($F_{1,17956}$ = 250.718; $P < 0.001$). There was also a significant interaction between domain and thermophily ($F_{1,17956} = 13.99$; $P < 0.001$). Consistent with previous results (Lao and Forsdyke 2000; Lambros *et al.* 2003), in both Bacteria and Archaea, mean AG3 was higher in thermophiles than in nonthermophiles (Figure 2b). This difference was more pronounced in Archaea than in Bacteria (Figure 2b).

In contrast to the case of GC3, there was a modest but significant positive correlation between $d_S$ and AG3 ($r$ = 0.090; $P < 0.001$). As with GC3, when AG3 was included as a covariate in the analysis of variance in $d_N/d_S$, there were significant effects of both the covariate ($F_{1,17956}$ = 25.21; $P < 0.001$) and thermophily ($F_{1,17956}$ = 24.31; $P < 0.001$). Thus, the association between thermophily and a reduced $d_N/d_S$ was statistically independent of the linear relationship between $d_N/d_S$ and AG3.

In certain comparisons, the ML method estimated that multiple substitutions had occurred per site, especially in the case of synonymous sites. Although some authors have argued that such estimates are reliable (Blanc *et al.* 2003), the accuracy of estimation may be problematic because the sites have been saturated with changes. To test whether our observed results were affected by saturation, we reanalyzed the data after excluding all comparisons for which $d_S$ or $d_N$ was estimated to be ≥1.5. The results (Figure 1b) were similar to those observed with the complete data set. In a factorial analysis of variance the only significant effect was that of thermophily ($F_{1,9966}$ = 7.49; $P = 0.006$). When GC3 was included in the model as a covariate, there was no significant effect of the covariate, but there was a significant effect of thermophily ($F_{1,9965}$ = 7.41; $P = 0.007$). Similarly, when AG3 was included as a covariate, there was no significant effect of the covariate, but there was a significant effect of thermophily ($F_{1,9965}$ = 7.61; $P = 0.006$).

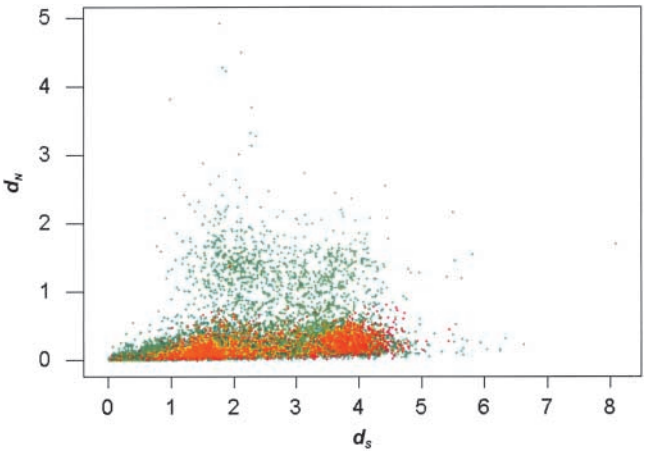


Figure 3.—Plot of $d_N$ *vs.* $d_S$ for comparisons of orthologous gene pairs from thermophilic (red dots) and nonthermophilic (green dots) species of Bacteria and Archaea. Analysis of covariance in $d_N$ with $d_S$ as covariate showed a significant effect of the covariate ($F_{1,17956}$ = 1072.16; $P < 0.001$), a significant difference between thermophiles and nonthermophiles ($F_{1,17956}$ = 7.15; $P = 0.007$), and a significant interaction between thermophily and the covariate ($F_{1,17956}$ = 133.60; $P < 0.001$).

In the complete data set, the correlation coefficient between $d_N$ and $d_S$ in thermophiles ($r = 0.485$; $P < 0.001$) was similar to that in nonthermophiles ($r$ = 0.454; $P < 0.001$). However, analysis of covariance indicated that there was a significant difference between thermophiles and nonthermophiles with respect to the slope of the linear relationship between $d_N$ and $d_S$ (Figure 3). We fitted linear regression lines through the origin on the assumption that $d_N$ and $d_S$ are both initially zero at the moment of lineage divergence for thermophiles and nonthermophiles. The slope of the former line was 0.077, while that of the latter was 0.129. These results suggest that $d_N$ increases for a given increase in $d_S$ ~1.67 times faster in nonthermophiles than in thermophiles.

Our results are consistent with the hypothesis that the proteins of thermophiles are subject to unusually strong functional constraints in amino acid sequence, which is reflected in a reduced level of nonsynonymous nucleotide substitution for a given level of synonymous substitution. As a further test of the hypothesis that this constraint arises from the high-temperature environment, we made pairwise comparisons of shared orthologs between the hyperthermophilic archaeal genus Pyrococcus and both the mesophilic genus Methanosarcina and the moderately thermophilic genus Sulfolobus (Madigan *et al.* 2003).

Examining pairs of orthologs present in both genera, we found that $d_N/d_S$ ratios for the same genes were significantly higher in Methanosarcina than in Pyrococcus (Figure 4a). The mean for Methanosarcina (0.074 ± 0.003 SEM) differed significantly from that for Pyrococ-

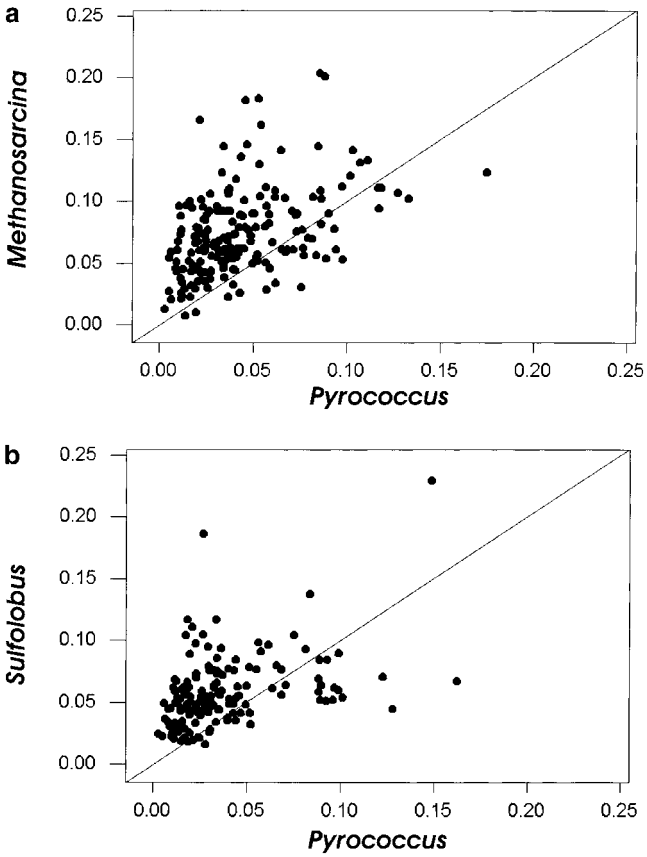


FIGURE 4.—Plots of $d_N/d_S$ for orthologous gene pairs shared by Archaeal genera. In each case the line has a slope of 1. (a) $d_N/d_S$ for *Methanosarcina acetivorans vs. M. mazei* plotted against $d_N/d_S$ for the orthologous genes in *Pyrococcus abyssi vs. P. horokoshii* ($N = 192$). (b) $d_N/d_S$ for *Sulfolobus solfataricus vs. S. tokodaii* plotted against $d_N/d_S$ for the orthologous genes in *P. abyssi vs. P. horokoshii* ($N = 152$).

cus ($0.044 \pm 0.002$; paired $t = 12.29$; two-tailed $P < 0.001$). Similarly, $d_N/d_S$ ratios were significantly higher in Sulfolobus than in Pyrococcus (Figure 4b). The mean for Methanosarcina ($0.0584 \pm 0.002$) differed significantly from that for Pyrococcus ($0.037 \pm 0.002$; paired $t = 7.94$; two-tailed $P < 0.001$). This result provided evidence of an unusually reduced rate of nonsynonymous substitution in a hyperthermophile. This, in turn, provides additional support for the hypothesis that life at high temperatures imposes unusual functional constraints on the primary structure of proteins.

## DISCUSSION

An analysis of mutation in a thermophilic archaeon (GROGAN *et al.* 2001) revealed a mutation rate close to or slightly less than that characteristic of all DNA-based microbes examined to date. However, the fraction of mutations that were base substitutions was half or less of that observed in eubacterial and eukaryotic microbes and in mammals. This observation generated the hypothesis that the average amino acid substitution is more deleterious in a thermophile than in a nonthermophile. On this hypothesis, the especially deleterious nature of mutation in thermophiles would have favored the evolutionary fixation of modifiers that specifically decrease the rates of base-substitution mutagenesis. Strong selection against amino acid replacements in thermophiles would impact their molecular evolution in two ways. First, because most selectively neutral mutations that reach fixation are base substitutions and not indels, the overall rate of molecular evolution would be reduced. Second, within this generally reduced rate of molecular evolution, the fixation of nonsynonymous substitutions would be reduced compared to that of synonymous substitutions. This study provides support for the latter prediction.

It has been suggested that Archaea evolve more slowly than either Bacteria or Eukaryotes. A recent test of this hypothesis concluded that amino acid sequences of Bacteria and Eukaryotes evolve at indistinguishable rates but that those of Archaea evolve 10–40% more slowly (KOLLMAN and DOOLITTLE 2000). However, in that analysis, 10% of the Bacteria but ~78% of the Archaea were thermophiles. By contrast, our results imply that natural selection eliminates a higher proportion of nonsynonymous mutations in thermophiles than in nonthermophiles, resulting in a lower average rate of amino acid replacement per mutational event in thermophiles, whereas nonthermophilic Archaea and nonthermophilic Bacteria are indistinguishable with respect to $d_N/d_S$. Likewise, thermophilic Archaea and thermophilic Bacteria are indistinguishable with respect to $d_N/d_S$. Thus, our results imply both that strong purifying selection against amino acid changes in thermophiles operates in a similar fashion across both prokaryotic domains and that rates of evolution are likely to be indistinguishable in Bacteria and Archaea but to depend critically on optimal growth temperature.

Examination of protein-coding sequences from a wide variety of organisms has provided evidence of the near ubiquity of "purifying" or conservative natural selection, which acts to eliminate deleterious mutations (NEI 1987). Because of their effect on protein structure, nonsynonymous mutations are more likely than synonymous mutations to be deleterious. As a result, over time the rate of synonymous nucleotide substitution per synonymous site ($d_S$) is predicted to exceed the rate of nonsynonymous nucleotide substitution per nonsynonymous site ($d_N$), and this prediction is supported in the case of most genes (KIMURA 1977; NEI 1987). There is evidence that some, if not most, prokaryote codon usage is subject to selective constraint (SHARP and LI 1986; BULMER 1991; LYNN *et al.* 2002). Where selection favors certain synonymous codons over others, purifying selection may occur at synonymous as well as nonsynonymous sites. However, that overall $d_N/d_S$ ratios in prokaryotes are $\ll 1.0$ (Figure 1) indicates that purifying selection

at synonymous sites is of negligible magnitude in comparison to that at nonsynonymous sites.

In a survey of 40 prokaryotic genomes, Lynn *et al.* (2002) found that the factor explaining the largest proportion (~25%) of the variance in codon usage was G + C content in the genome. The second most important factor, related to optimal growth temperature, accounted for an additional 10% of the variance (Lynn *et al.* 2002). These authors found that the major difference between thermophiles and nonthermophiles related to their use of arginine and isoleucine codons. Other studies have suggested that thermophiles have a preference for purines in mRNAs, which is also expected to affect codon usage (Lao and Forsdyke 2000; Lambros *et al.* 2003). If thermophiles are subject to greater constraint on codon usage than nonthermophiles, this would be expected to reduce $d_S$ and thus to increase the $d_N/d_S$ ratio in thermophiles. That we observed overall lower mean $d_N/d_S$ in thermophiles than in nonthermophiles suggests that, even if greater constraint on codon usage is present in thermophiles than in nonthermophiles, such constraint is not of sufficient magnitude to mask the strong difference with respect to constraint on amino acid sequences.

The factors that influence variations among organisms with respect to patterns of synonymous and nonsynonymous nucleotide substitution remain poorly understood. Our results suggest that one such factor can be genome-wide differences among organisms with respect to the strength of purifying selection. Furthermore, our results demonstrate that a species' environment represents a source of selection that can leave a clear imprint on the pattern of long-term genomic evolution.

We are grateful for comments on the manuscript by D. W. Grogan. This research was supported by grant GM66710 to A.L.H. from the National Institutes of Health.

## LITERATURE CITED

Altschul, S. F., T. L. Madden, A. A. Schaffer, J. Zhang, Z. Zhang *et al.*, 1997 Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. **25:** 3389–3402.

Blanc, G., K. Hokamp and K. H. Wolfe, 2003 A recent polyploidy superimposed on older large-scale duplications in the *Arabidopsis* genome. Genome Res. **13:** 137–144.

Brown, J. R., and A. N. Lupas, 1998 What makes a thermophile? Trends Microbiol. **6:** 349–351.

Bulmer, M., 1991 The selection-mutation-drift theory of synonymous codon usage. Genetics **129:** 897–907.

Cambillau, C., and J. M. Claverie, 2000 Structural and genomic correlates of hyperthermostability. J. Biol. Chem. **275:** 32383–32386.

Chen, Y. W., M. Bycroft and K.-B. Wong, 2003 Crystal structure of ribosomal protein L30e from the extreme thermophile *Thermococcus celer*: thermal stability and RNA binding. Biochemistry **42:** 2857–2865.

Criswell, A. R., E. Bae, B. Stec, J. Konisky and G. N. Phillips, Jr., 2003 Structures of thermophilic and mesophilic adenylate kinases from the genus *Methanococcus*. J. Mol. Biol. **333:** 1087–1099.

Felsenstein, J., 1985 Phylogenies and the comparative method. Am. Nat. **125:** 1–15.

Grogan, D. W., G. T. Carver and J. W. Drake, 2001 Genetic fidelity under harsh conditions: analysis of spontaneous mutation in the thermoacidophilic archaeon *Sulfolobus acidocaldarius*. Proc. Natl. Acad. Sci. USA **98:** 7928–7933.

Gromiha, M. M., M. Oobatake and A. Sarai, 1999 Important amino acid properties for enhanced thermostability from mesophilic to thermophilic proteins. Biophys. Chem. **82:** 51–67.

Kimura, M., 1977 Preponderance of synonymous changes as evidence for the neutral theory of molecular evolution. Nature **267:** 275–276.

Kollman, J. M., and R. F. Doolittle, 2000 Determination of the relative rates of change for prokaryotic and eukaryotic proteins with anciently duplicated paralogs. J. Mol. Evol. **51:** 173–181.

Kreil, D. P., and C. A. Ouzounis, 2001 Identification of thermophilic species by the amino acid compositions deduced from their genomes. Nucleic Acids Res. **29:** 1608–1615.

Kumar, S., and R. Nussinov, 2001 How do thermophilic proteins deal with heat? Cell. Mol. Life Sci. **58:** 1216–1233.

Kumar, S., C. J. Tsai and R. Nussinov, 2000 Factors enhancing protein thermostability. Protein Eng. **13:** 179–191.

Lambros, R. J., J. R. Mortimer and D. R. Forsdyke, 2003 Optimum growth temperature and the base composition of open reading frames in prokaryotes. Extremophiles **7:** 443–450.

Lao, P. J., and D. R. Forsdyke, 2000 Thermophilic bacteria strictly obey Szybalski's transcription direction rule and politely purine-load RNAs with both adenine and guanine. Genome Res. **10:** 228–236.

Lynn, D. J., G. A. C. Singer and D. A. Hickey, 2002 Synonymous codon usage is subject to selection in thermophilic bacteria. Nucleic Acids Res. **30:** 4272–4277.

Madigan, M. T., J. M. Martinko and J. Parker, 2003 *Brock Biology of Microorganisms*, Ed. 10. Prentice Hall, Upper Saddle River, NJ.

Nei, M., 1987 *Molecular Evolutionary Genetics*. Columbia University Press, New York.

Perutz, M. F., 1978 Electrostatic effects in proteins. Science **201:** 1187–1191.

Sharp, P. M., and W.-H. Li, 1986 An evolutionary perspective on synonymous codon usage in unicellular organisms. J. Mol. Evol. **24:** 28–38.

Tekaia, F., E. Yeramian and B. Dujon, 2002 Amino acid composition of genomes, lifestyles of organisms, and evolutionary trends: a global picture with correspondence analysis. Gene **297:** 51–60.

Thompson, J. D., D. G. Higgins and T. J. Gibson, 1994 CLUSTAL W: improvement of the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res. **22:** 4673–4680.

Woese, C. R., O. Kandler and M. L. Wheelis, 1990 Towards a natural system of organisms: proposal for the domains Archaea, Bacteria, and Eucarya. Proc. Natl. Acad. Sci. USA **87:** 4576–4579.

Wolfe, K. H., P. M. Sharp and W.-H. Li, 1989 Mutation rates differ among regions of the mammalian genome. Nature **337:** 283–285.

Yang, Z., 1997 PAML: a program package for phylogenetic analysis by maximum likelihood. Comput. Appl. Biosci. **13:** 555–556.

Yang, Z., and R. Nielsen, 2000 Estimating synonymous and nonsynonymous substitution rates under realistic evolutionary models. Mol. Biol. Evol. **17:** 32–43.

Zhang, J., H. F. Rosenberg and M. Nei, 1998 Positive Darwinian selection after gene duplication in primate ribonuclease genes. Proc. Natl. Acad. Sci. USA **98:** 3708–3713.

Communicating editor: S. W. Schaeffer