

Nucleotide Bias Causes a Genomewide Bias in the Amino Acid Composition of Proteins

Gregory A. C. Singer and Donal A. Hickey

Department of Biology, University of Ottawa, Ottawa, Ontario, Canada

We analyzed the nucleotide contents of several completely sequenced genomes, and we show that nucleotide bias can have a dramatic effect on the amino acid composition of the encoded proteins. By surveying the genes in 21 completely sequenced eubacterial and archaeal genomes, along with the entire *Saccharomyces cerevisiae* genome and two *Plasmodium falciparum* chromosomes, we show that biased DNA encodes biased proteins on a genomewide scale. The predicted bias affects virtually all genes within the genome, and it could be clearly seen even when we limited the analysis to sets of homologous gene sequences. Parallel patterns of compositional bias were found within the archaea and the eubacteria. We also found a positive correlation between the degree of amino acid bias and the magnitude of protein sequence divergence. We conclude that mutational bias can have a major effect on the molecular evolution of proteins. These results could have important implications for the interpretation of protein-based molecular phylogenies and for the inference of functional protein adaptation from comparative sequence data.

Introduction

It is well known that both individual genes and entire genomes can vary significantly in nucleotide composition (Bernardi and Bernardi 1986; Muto and Osawa 1987). Some organisms, for example, have genomes that are disproportionately rich in guanine and cytosine (G and C), while others have DNA that is rich in adenine and thymine (A and T). Variation in nucleotide composition is usually most pronounced at the synonymous codon positions of genes, and, because of the redundancy in the genetic code, these variations in DNA content may have little effect on the amino acid content of the encoded proteins (Loomis and Smith 1990; Lockhart et al. 1992). If, however, compositional bias at the DNA level affects both the synonymous and the nonsynonymous sites in protein-coding genes, then we expect proteins to change their amino acid composition over evolutionary time, and in a direction predicted by the underlying nucleotide bias. Several previous studies suggest that protein evolution is affected by the nucleotide composition of the encoding genes. The first evidence was provided many years ago by the finding that there was a correlation between the nucleotide content and the amino acid content of bacterial cells (Sueoka 1961). Subsequently, a number of surveys of molecular sequences (Bernardi and Bernardi 1986; D'Onofrio et al. 1991; Collins and Jukes 1993; Berkhout and van Hermet 1994; Hickey, Wang, and Magoulas 1994; Porter 1995) identified correlations between the nucleotide composition of DNA and the amino acid content of the encoded proteins. Many of these earlier surveys, however, involved intragenomic comparisons of heterologous DNA and protein sequences. More recent studies have focused primarily on intergenomic comparisons of homologous gene sequences (Foster, Jermin, and Hickey 1997; Li 1997, pp. 422–423; Lobry 1997; Gu, Hew-

ett-Emmett, and Li 1998; D'Onofrio et al. 1999; Lafay et al. 1999; Rodriguez-Trelles, Tarrío, and Ayala 1999; Wilquet and Van de Castele 1999). All of these studies, despite the fact that they were based on a wide variety of different genes and proteins, provide evidence for a significant correlation between DNA composition and protein composition. We wished to test the generality of these correlations through a comprehensive analysis of completely sequenced genomes.

There are currently more than 20 completely sequenced prokaryotic genomes and 2 completely sequenced eukaryotic genomes available in the public databases. One of the most striking features of these genomes is the range of nucleotide compositions represented. For example, *Borrelia burgdorferi* has an overall GC content of only 25.5% in its coding DNA, while *Mycobacterium tuberculosis* has a GC content of 65.9%. With such a range in nucleotide composition, these genomes provide an excellent resource for studying the relationship between DNA bias and amino acid composition. We decided to perform a predictive analysis similar to that used by Foster, Jermin, and Hickey (1997): by partitioning the universal codon table into GC-rich codons, AT-rich codons, and neutral codons (see table 1), we could make the simple prediction that AT-rich coding sequences would encode proteins rich in the FYMINK amino acids (phenylalanine, tyrosine, methionine, isoleucine, asparagine, and lysine), while GC-rich coding sequences would produce proteins containing high levels of the GARP amino acids (glycine, alanine, arginine, and proline). In our comprehensive analysis, involving 22 completely sequenced genomes, we have shown that this prediction holds true. Moreover, the prediction holds true for every gene examined in highly biased genomes.

Materials and Methods

Twenty-one prokaryotic genomes, the entire *Saccharomyces cerevisiae* genome, and two *Plasmodium falciparum* chromosomes were downloaded from the GenBank FTP site (<ftp://ftp.ncbi.nlm.nih.gov/genbank/genomes/>): *S. cerevisiae* (U00091, Y13134, X59720,

Key words: nucleotide content, amino acid composition, proteome composition, genomes, codon usage, proteomics, bioinformatics.

Address for correspondence and reprints: Donal A. Hickey, Department of Biology, University of Ottawa, 30 Marie Curie Street, Ottawa, Ontario, Canada K1N 6N5. E-mail: dhickey@uottawa.ca.

Mol. Biol. Evol. 17(11):1581–1588. 2000

© 2000 by the Society for Molecular Biology and Evolution. ISSN: 0737-4038

Table 1
The Universal Codon Table, Rearranged to Group Codons According to Their Nucleotide Compositions

FIRST CODON POSITION	SECOND CODON POSITION							
	A or T			G or C				
A or T ...	AAG	Lys	ATG	Met	AGG	Arg	ACG	Thr
	AAA		ATA		AGA		ACA	
	AAC	Asn	ATC	Ile	AGC	Ser	ACC	Ser
	AAT		ATT		AGT		ACT	
	TAG	STP	TTG	Leu ^a	TGG	Trp ^b	TCG	Ala
	TAA		TTA		TGA		TCA	
	TAC	Tyr	TTC	Phe	TGC	Cys	TCC	Pro
	TAT		TTT		TGT		TCT	
G or C ...	GAG	Glu	GTG	Val	GGG	Gly	GCG	Ala
	GAA		GTA		GGA		GCA	
	GAC	Asp	GTC		GGC		GCC	Pro
	GAT		GTT		GGT		GCT	
	CAG	Gln	CTG		CGG		CCG	Pro
	CAA		CTA	Leu	CGA	Arg ^c	CCA	
	CAC	His	CTC		CGC		CCC	Pro
	CAT		CTT		CGT		CCT	

^a Because four of the six leucine codons are not AT-rich, leucine was not included in the AT-rich amino acid group.

^b While tryptophan has a GC-rich codon, the *Mycoplasma* genomes also use TGA to encode this amino acid, so its codons are not consistently GC-rich. We therefore did not include tryptophan in our group of GC-rich amino acids.

^c Arginine was included in our GC-rich amino acid group because the majority of its codons (4/6) are GC-rich.

Z71256, U00092, D50617, Y13135, U00093, Z47047, Y13136, Y13137, Y13138, Z71257, Y13139, Y13140, U00094 for chromosomes 1–16, respectively; 6,274 genes; 40.3% GC in the coding sequences); *P. falciparum* (AE001362, MAL3 for chromosomes 2 and 3, respectively; 421 genes; 23.4% GC in the coding sequences); *Aeropyrum pernix* K1 (Aero.p; 2,694 genes; 57.5% GC); *Archaeoglobus fulgidus* (AE000782; 2,407 genes; 49.4% GC); *Aquifex aeolicus* (AE000657; 1,522 genes; 43.7% GC); *B. burgdorferi* (AE000783 and seq-AE000784–seq-AE000794 for 11 plasmids; 1,254 genes; 25.5% GC); *Bacillus subtilis* (AL009126; 4,100 genes; 44.3% GC); *Chlamydia pneumoniae* (AE001363; 1,052 genes; 41.3% GC); *Chlamydia trachomatis* (AE001273; 894 genes; 41.7% GC); *Escherichia coli* K-12 MG1655 (U00096; 4,289 genes; 51.8% GC); *Haemophilus influenzae* Rd (L42023; 1,709 genes; 38.8% GC); *Helicobacter pylori* 26695 (AE000511; 1,566 genes; 39.6% GC); *Helicobacter pylori* J99 (AE001439; 1,491 genes; 39.9% GC); *Mycoplasma genitalium* G37 (L43967; 480 genes; 31.6% GC); *Methanococcus jannaschii* (L77117; 1,715 genes; 31.9% GC); *Mycoplasma pneumoniae* M129 (U00089; 677 genes; 40.7% GC); *Methanobacterium thermoautotrophicum* delta H (AE000666; 1,869 genes; 50.6% GC); *Mycobacterium tuberculosis* (AL123456; 3,918 genes; 65.9% GC); *Pyrococcus horikoshii* OT3 (Pyro.h; 2,064 genes; 42.3% GC); *Rickettsia prowazekii* Madrid E (AJ235269; 834 genes; 30.4% GC); *Synechocystis* PCC6803 (AB001339; 3,169 genes; 48.6% GC); *Thermotoga maritima* (AE000512; 1,846 genes; 46.4% GC); *Treponema pallidum* (AE000520; 1,031 genes; 52.6% GC). Both nucleotide and amino acid sequences were extracted directly from the GenBank flat files.

In the manner introduced by Foster, Jermin, and Hickey (1997), we partitioned the codon table into three groups: codons that were GC-rich at the first two codon positions, those that were AT-rich at the first two codon positions, and unbiased codons (table 1). The AT-rich codons encoded the FYMINK amino acids (phenylalanine, tyrosine, methionine, isoleucine, asparagine, and lysine). Although there were two AT-rich leucine codons, leucine was not included among this set because there were four other leucine codons that were not AT-rich. The GC-rich codons encoded the GARP amino acids (glycine, alanine, arginine, and proline). In this set, arginine is included because four of its six codons are GC-rich.

Homologous genes common to all genomes were identified by performing gapped-BLASTP searches (Altschul et al. 1997) of each genome against *B. subtilis*, a large and relatively unbiased genome. A cutoff expect score of 0.0001 was used, and in addition, hits with less than 25% identities were ignored; this second criterion resulted in the removal of many of the lower BLAST scores. Hits common to all BLAST searches were assumed to be homologs. In the comparison of homologous gene pairs between *B. burgdorferi* and *M. tuberculosis*, the median expect score was 9×10^{-39} , and the minimum expect score was 9×10^{-5} .

Results

Comparison of Complete Genomes

In our first analysis, we included all of the genes and all of the protein sequences from these genomes. We predicted that AT-rich (GC-poor) genomes would have a higher proportion of AT-rich codons in their genes, which encode the FYMINK amino acids. Likewise, GC-rich genomes should have proteins that are rich in the GARP amino acids. Our results showed that this relationship holds true for the genomes analyzed in this study (fig. 1A and B), and these results are statistically highly significant ($P < 10^{-9}$).

In order to demonstrate a causal relationship between nucleotide bias and amino acid composition, as opposed to a simple correlation between the two, we wished to eliminate the alternative possibility, i.e., that amino acid biases were affecting the nucleotide composition of genes rather than vice versa. This was accomplished by repeating the analysis using the nucleotide composition at the synonymous codon positions only. These results are shown in figure 1C and D and show that there is still a very significant correlation between nucleotide composition and amino acid composition ($P < 10^{-6}$). Since the synonymous sites do not affect the protein sequence, this correlation reflects an underlying nucleotide bias affecting both synonymous and nonsynonymous sites. Additional supporting evidence for this conclusion stems from a comparison of the relative extents of the nucleotide bias at synonymous versus nonsynonymous codon positions (fig. 2). Nucleotide bias is much greater at the synonymous codon positions than at the nonsynonymous positions. If there were a primary amino acid bias within proteins, which

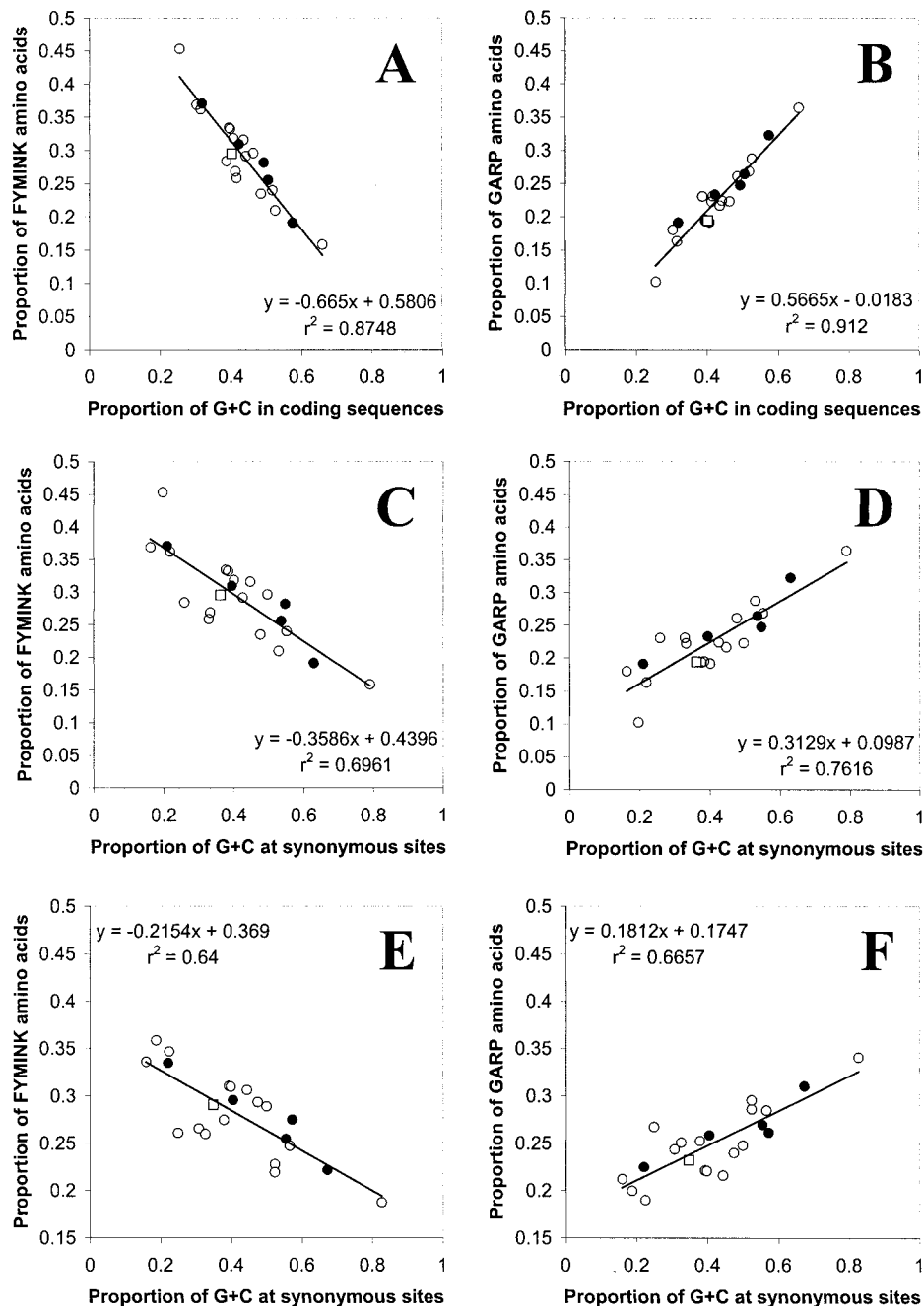


FIG. 1.—Correlation between mean nucleotide content and amino acid content of all encoded proteins for 22 complete genomes (see *Materials and Methods*). Both nucleotide and amino acid sequences for all genes in each genome were extracted from the GenBank flat files. Eubacterial genomes are shown as open circles, archaeal genomes are shown as closed circles, and the eukaryotic genome (*Saccharomyces cerevisiae*) is shown as a square. Because of the large sample sizes, the standard errors of the means were very small (usually less than 0.001); standard error bars are not shown in the figure, but if they were shown, they would be smaller than the diameter of the symbols indicating the means. *A*, The total genomic content of FYMINK amino acids (see text) decreases with increasing G+C content in the coding sequences. *B*, A corresponding increase in the proportion of GARP amino acids with increasing G+C content of the genomes. *C* and *D*, The results of a repeated analysis based on G+C content at the synonymous codon positions only. *E* and *F*, The relationship between amino acid content and synonymous-site nucleotide content for a core set of homologous genes that are shared by all of the genomes.

was then reflected as a secondary DNA bias, the nonsynonymous codon positions should be more biased than the synonymous codon positions, which are unaffected by amino acid content. Conversely, a primary bias at the DNA level is expected to be most pronounced at the synonymous codon positions, but less so at the

nonsynonymous codon positions, which are under stronger selective constraint. The data presented in figure 2 strongly support the latter interpretation. Thus, these results are consistent with the view that the differences in amino acid composition are a consequence, rather than a cause, of the nucleotide bias.

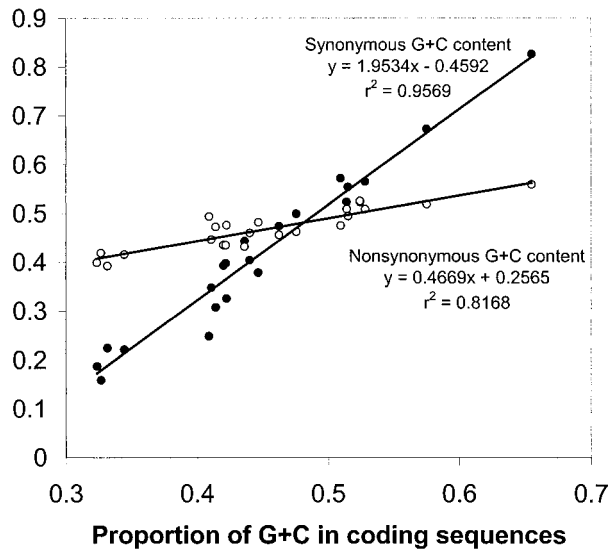


FIG. 2.—Comparison of nucleotide compositions at synonymous and nonsynonymous sites. For each of the genomes shown in figure 1, the G+C content at synonymous sites (filled circles) and nonsynonymous sites (open circles) were plotted against the total G+C content of the coding sequences.

Finally, we were concerned that our results might be skewed by the fact that not all genomes contain the same set of genes. For instance, some genomes might have a higher proportion of membrane-spanning proteins, and this could result in differences in amino acid composition that are functionally based. In order to eliminate this possibility, we performed an additional comparison, using only the amino acid and nucleotide compositions of genes that are shared between these genomes. In this way, we could separate the effects of nucleotide bias from the effects of functional constraint. Homologous genes were identified by sequence similarity searches. Although we found hundreds of genes in common between any pair of genomes (see below), we limited this analysis to a core set of 47 genes that were common to all 22 genomes. Again, we found a strong relationship ($P < 10^{-10}$) between the nucleotide content at synonymous sites and amino acid composition for this set of homologous genes (fig. 1E and F). It is important to note that most of the variability in amino acid frequencies between these sequences can be explained by nucleotide bias ($r^2 = 0.64$ for the FYMINK amino acids, and 0.67 for the GARP amino acids). Even more surprising is that because these genes were identified via stringent sequence similarity searches, only the most conserved genes were included in the analysis. Since we have shown that rapidly evolving sequences are more prone to nucleotide bias (see below), the results based on this relatively conserved group of genes gives an underestimate of the average effect of nucleotide bias on amino acid composition. In fact, this restriction of the analysis to relatively conserved sequences explains why the slopes of the regression lines in figure 1E and F are less than those in figure 1C and D.

Because phylogenetically related species are expected to have similar nucleotide and amino acid biases,

it is plausible that the results shown in figure 1A–F reflect phylogenetic history rather than mutational bias. Pagel and Harvey (1988) have summarized the primary methods to distinguish evolutionary trends from phylogenetic artifacts. All of these methods depend on some degree of replication of the trend within phylogenetically independent lineages. With our data set, we had an opportunity to look for repeated trends within two distinct lineages, the eubacteria and the archaea. A simple inspection of the distribution of open circles (eubacteria) and filled circles (archaea) in figure 1A–F suggests that the trend is indeed repeated within both of these two major lineages and that no significant difference exists between these two sets of data. In order to verify this impression, new regressions were calculated for each of these two kingdoms. We found that the regression was significant ($P < 0.003$) within both the eubacteria and the archaea, indicating that the trend is indeed repeated. In addition, our comparison of *Plasmodium* and *Saccharomyces* proteins (see below) indicated that a similar trend was also repeated within the eukaryotic kingdom. We then used the analysis of covariance to test for differences between the slopes of the regression lines for the eubacterial and the archaeal data sets. We found no significant difference between the two groups. In other words, not only is the trend repeated, but also the magnitude of the effect is very similar within both the eubacteria and the archaea. This is entirely consistent with the impression obtained from the simple visual inspection of the data points in figure 1.

Amino Acid Changes Within Individual Genes

In order to examine the effects of nucleotide bias on individual genes in more detail, we decided to compare the two genomes with the most distinct nucleotide compositions available. We identified the genes in common between the genomes of the highly AT-rich genome of *B. burgdorferi* and the very GC-rich genome of *M. tuberculosis* in an analysis similar to that used by Lafay et al. (1999). We identified 305 genes common to these two organisms, and we measured the synonymous nucleotide frequencies and amino acid contents of each one (fig. 3A and B). Although there was no overlap in the synonymous GC contents of these two genomes, there did appear to be some overlap in the amino acid proportions of the encoded proteins. This apparent overlap disappeared, however, when we compared homologous gene pairs (fig. 4). For every gene, the GARP/FYMINK ratio in the *M. tuberculosis* homolog was higher than that of the corresponding gene in *B. burgdorferi* (fig. 4). A one-tailed paired-sample *t*-test showed this difference to be highly significant ($P = 3 \times 10^{-122}$). In effect, these results can be viewed as a replicated experiment. Each gene underwent two different evolutionary “treatments”: one evolved in an AT-biased environment, while the other evolved in a GC-rich environment. This “experiment” was repeated for 305 genes, and all of them evolved in the directions we predicted.

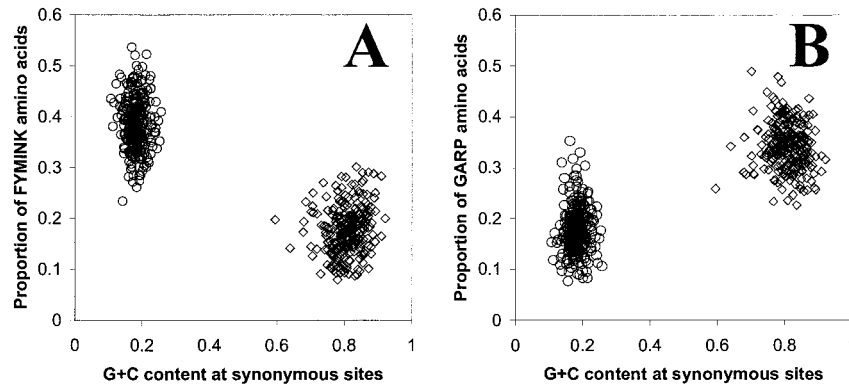


FIG. 3.—Amino acid content of homologous genes in *Borrelia burgdorferi* and *Mycobacterium tuberculosis*. A total of 305 homologous genes were identified by BLAST searches. Each point represents the data for a single gene. The graphs show the relationship between the amino acid content of each gene and the G+C content at the synonymous codon positions of that gene. Data for *B. burgdorferi* genes are shown as circles, and data points for *M. tuberculosis* genes are shown as diamonds. For further explanation, see figure 1.

To compare the effects of DNA mutation bias on highly conserved genes versus weakly conserved genes, we plotted the protein composition against the sequence divergence between the homologous gene pairs (fig. 5). There is highly significant correlation between the degree of sequence divergence and the magnitude of the compositional bias ($P < 10^{-30}$). In table 2, we show representatives of both conserved and diverged gene pairs. The results show, as one might expect, that the more weakly conserved *tylA* and *rnhB* genes underwent striking changes in their amino acid compositions (table 2). Remarkably, even the strongly conserved *gap* and *tuf* genes show amino acid composition changes in the predicted directions, although to a more moderate degree than the more rapidly diverging genes.

Examination of Individual Amino Acid Changes

Our prediction was that GC-rich coding DNA would produce proteins rich in the GARP amino acids,

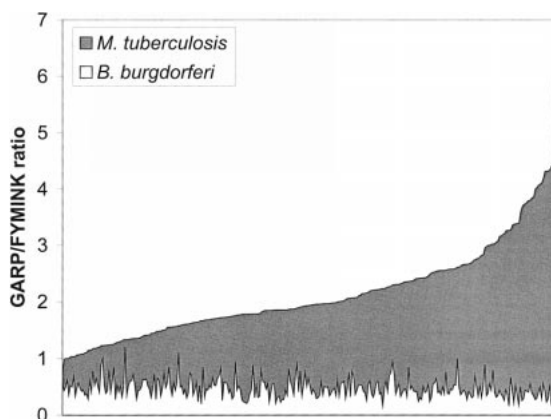


FIG. 4.—Pairwise comparison of homologous genes from *Mycobacterium tuberculosis* and *Borrelia burgdorferi*. The same 305 genes shown in figure 3 were used for this analysis. First, the 305 *M. tuberculosis* genes are arranged in order of increasing G+C content on the horizontal axis, and the GARP/FYMINK ratio of the encoded proteins are plotted on the vertical axis (shaded area). For each homologous gene from *B. burgdorferi*, the same ratio is plotted (white area). All 305 genes analyzed showed higher GARP/FYMINK ratios in the GC-rich *M. tuberculosis* than in the AT-rich *B. burgdorferi*.

while AT-rich DNA would produce FYMINK-rich proteins. In general, we have shown that this is the case. Nevertheless, we decided that it was important to determine whether certain amino acids within each group were more apt to change than others. We scored the proportions of individual amino acids within the 305 gene pairs of the AT-rich *B. burgdorferi* and the GC-rich *M. tuberculosis*. The results are shown in figure 6. The GARP and FYMINK amino acids all changed proportions in the predicted directions. Because of the number of sites involved (and the resulting power of the statistical test), every amino acid proportion was found to be highly statistically different between the two organisms ($P < 10^{-24}$ in a one-tailed paired-sample *t*-test), with the exceptions of methionine ($P = 0.36$), cysteine ($P = 0.006$), and leucine ($P = 0.017$). We had predicted

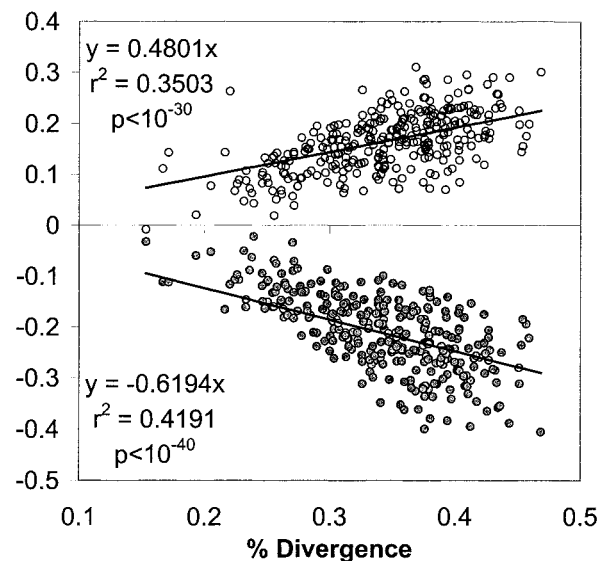


FIG. 5.—The relationship between the degree of amino acid bias and sequence divergence. For each homologous gene pair in the genomes of *Borrelia burgdorferi* and *Mycobacterium tuberculosis*, the increase in the proportion of GARP amino acids (open circles) and the decrease in the proportion of FYMINK amino acids (shaded circles) were plotted for the *M. tuberculosis* gene relative to the *B. burgdorferi* homolog.

Table 2
A Comparison of the Proportions of FYMINK and GARP Amino Acids in Four Homologous Gene Pairs from the Genomes of *Borrelia burgdorferi* and *Mycobacterium tuberculosis*

GENE ^a	<i>B. BURGENDORFERI</i>		<i>M. TUBERCULOSIS</i>	
	FYMINK	GARP	FYMINK	GARP
Hemolysin (<i>tlyA</i>)	108 (40.9%)	31 (11.7%)	24 (8.9%)	108 (40.1%)
Ribonuclease H (<i>rnhB</i>)	81 (44.5%)	30 (16.5%)	36 (13.6%)	114 (43.0%)
Glyceraldehyde 3-phosphate dehydrogenase (<i>gap</i>)	91 (27.1%)	77 (22.9%)	74 (21.8%)	94 (27.6%)
Translation elongation factor TU (<i>tuf</i>)	113 (28.1%)	101 (25.1%)	87 (21.9%)	108 (27.2%)

^a *tlyA* and *rnhB* are loosely conserved genes, while *gap* and *tuf* are very highly conserved.

that the proportion of methionine residues would increase in the AT-rich genome, and, while this is true, the observed effect is relatively slight. On closer inspection of the data, we observed that while methionines replace other more GC-rich codons in this genome, the methionine codon, ATG, is itself replaced by the even more AT-rich isoleucine codons, ATA and ATT. Likewise, the relatively high proportion of valine in *M. tuberculosis*, which is not predicted by our simple model, may be explained by inferring a single transition from A to G in the first position of the AT-rich isoleucine codons. This change produces a more GC-rich codon that encodes the biochemically similar amino acid, valine.

Similar Trends in Eubacterial, Archaeal, and Eukaryotic Genomes

When we plotted the relationship between nucleotide bias and amino acid content for the entire set of genomes examined (see fig. 1), we were surprised to see that there was no “clumping” of major phylogenetic groups in these graphs and that the archaeal and eubacterial genomes behaved as a single homogenous data set. Moreover, the yeast genome data also fell at the predicted point on these graphs. This suggests that the effects of nucleotide bias on protein composition are operating in all major lineages.

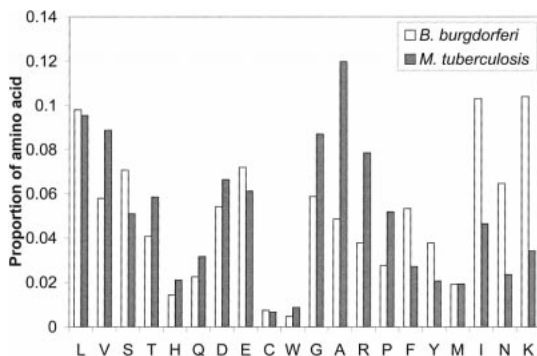


FIG. 6.—Changes in the proportions of individual amino acids between homologous protein sequences. The data are based on the set of 305 homologous genes in the genomes of *Borrelia burgdorferi* (white bars) and *Mycobacterium tuberculosis* (shaded bars). Amino acid and nucleotide frequencies were calculated for concatenated nucleotide and amino acid sequences. All of the differences are highly significant ($P < 10^{-35}$ in a one-tailed paired-sample *t*-test), with the exceptions of leucine (L), cysteine (C), and methionine (M).

Very few completely sequenced eukaryotic genomes are available yet, but we compared chromosomes 2 and 3 from the extremely AT-rich genome of *P. falciparum* with the entire *S. cerevisiae* genome, which is relatively unbiased, to see if the amino acid usage of these organisms was also affected by nucleotide bias. We found 90 genes common to these two genomes, and we measured the GARP/FYMINK ratio for each gene. Our findings show that, with the exception of a single gene (the 40S ribosomal protein), the relatively unbiased *S. cerevisiae* genes have a greater GARP/FYMINK ratio than the corresponding genes in *P. falciparum* (fig. 7), and this result is also highly significant ($P = 4.6 \times 10^{-33}$ in a one-tailed paired-sample *t*-test).

Discussion

Our main finding is not just that protein composition is affected by nucleotide bias, but also that this effect is both very large and very widespread. We were surprised that our very simple predictive model based on the genetic code table could explain so much of the intergenomic variation in protein composition. This result has implications for many studies that are based on the interpretation of amino acid sequence data. For instance, it has already been shown that nucleotide bias can affect the functional properties of proteins (Naylor,

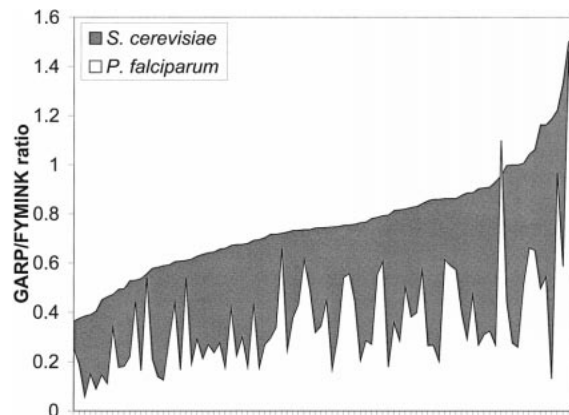


FIG. 7.—Effect of nucleotide bias on protein composition in eukaryotic genomes. Ninety homologous genes were identified in the genomes of *Plasmodium falciparum* (white area) and *Saccharomyces cerevisiae* (shaded area). The data were analyzed as described in figure 4. All but one of the homologous proteins show a higher GARP/FYMINK ratio in the GC-neutral *S. cerevisiae* than in the AT-rich *P. falciparum*.

Collins, and Brown 1995; Gu, Hewett-Emmett, and Li 1998; D'Onofrio et al. 1999) and that convergent amino acid composition can affect the construction of phylogenetic trees based on protein sequences (Gu, Hewett-Emmett, and Li 1998; Foster and Hickey 1999). For instance, Foster and Hickey (1999) showed that unrelated taxa were grouped together in a phylogenetic tree due to convergent amino acid sequences. Although this problem in phylogenetic reconstruction has been identified, a satisfactory method of dealing with the problem has not yet been found. Because of the relationship between primary amino acid sequence and secondary protein structure, nucleotide bias may also affect the evolution of protein structure (Wood and Pearson 1999).

The most parsimonious explanation of the observed patterns of amino acid composition in these genomes is an underlying mutational bias that varies between lineages (Sueoka 1988). The resulting amino acid sequence changes are nonrandom, since the mutational bias is strongly directional, and yet they are not caused by natural selection acting directly on protein function. Consequently, their evolutionary dynamics cannot be described in terms of either Darwinian selection or random genetic drift. They may, however, result in secondary selective changes in the protein sequence. For example, amino acid bias could result in a change of the charge distribution within a protein, as well as an alteration of the protein's secondary and tertiary structures. Such proteins may then undergo positive selection at other sites to counter the potentially deleterious effects of these nucleotide bias-induced changes. The long-term result might be a cascade of compensatory changes to reduce the impact of amino acid bias on protein structure and function. The problem of distinguishing between functional constraint in protein sequences and mutation-driven biases in the composition of these same sequences will provide a future challenge for molecular evolutionists.

In conclusion, we recognize that other factors, such as selective constraint, adaptive change, and genetic drift, all play important roles in protein sequence evolution. The results presented here, however, demonstrate that mutational pressure on DNA composition can also be a very powerful and pervasive force in long-term protein evolution.

Acknowledgments

This work was supported by a research grant from the Natural Sciences and Engineering Research Council (NSERC) to D.A.H. We also thank Robert Charlebois, whose NeuroGadgets web site (<http://www.neurogadgets.com>) was a source of inspiration for this research.

LITERATURE CITED

- ALTSCHUL, S. F., T. L. MADDEN, A. A. SCHAFFER, J. ZHANG, Z. ZHANG, W. MILLER, and D. J. LIPMAN. 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* **25**:3389–3402.
- BERKHOUT, B., and F. J. VAN HERMERT. 1994. The unusual nucleotide content of the HIV RNA genome results in a biased amino acid composition of HIV proteins. *Nucleic Acids Res.* **22**:1705–1711.
- BERNARDI, G., and G. BERNARDI. 1986. Compositional constraints and genome evolution. *J. Mol. Evol.* **24**:1–11.
- COLLINS, D. W., and T. H. JUKES. 1993. Relationship between G+C in silent sites of codons and amino acid composition of human proteins. *J. Mol. Evol.* **36**:201–213.
- 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.
- D'ONOFRIO, G., D. MOUCHIROUD, B. AÏSSANI, C. GAUTER, and G. BERNARDI. 1991. Correlations between the compositional properties of human genes, codon usage, and amino acid composition of proteins. *J. Mol. Evol.* **32**:504–510.
- FOSTER, P. G., and D. A. HICKEY. 1999. Compositional bias may affect both DNA-based and protein-based phylogenetics reconstructions. *J. Mol. Evol.* **48**:284–290.
- FOSTER, P. G., L. S. JERMIIN, and D. A. HICKEY. 1997. Nucleotide composition bias affects amino acid content in proteins coded by animal mitochondria. *J. Mol. Evol.* **44**:282–288.
- 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.
- HICKEY, D. A., S. WANG, and C. MAGOULAS. 1994. Gene duplication, gene conversion and codon bias. Pp. 199–207 in B. GOLDING, ed. *Non-neutral evolution*. Chapman and Hall, New York.
- LAFAY, B., A. T. LLOYD, M. J. MCLEAN, K. M. DEVINE, P. M. SHARP, and K. H. WOLFE. 1999. Proteome composition and codon usage in spirochaetes: species-specific and DNA strand-specific mutational biases. *Nucleic Acids Res.* **27**:1642–1649.
- LI, W.-H. 1997. *Molecular evolution*. Sinauer, Sunderland, Mass.
- LOBRY, J. R. 1997. Influence of genomic G+C content on average amino-acid composition of proteins from 59 bacterial species. *Gene* **205**:309–316.
- LOCKHART, P. J., C. J. HOWE, D. A. BRYANT, T. J. BEANLAND, and A. W. D. LARKUM. 1992. Substitutional bias confounds inference of cyanelle origins from sequence data. *J. Mol. Evol.* **34**:153–162.
- LOOMIS, W. F., and D. W. SMITH. 1990. Molecular phylogeny of *Dictyostelium discoideum* by protein sequence comparison. *Proc. Natl. Acad. Sci. USA* **87**:9093–9097.
- MUTO, A., and S. OSAWA. 1987. The guanine and cytosine content of genomic DNA and bacterial evolution. *Proc. Natl. Acad. Sci. USA* **84**:166–169.
- NAYLOR, G. J. P., T. M. COLLINS, and W. M. BROWN. 1995. Hydrophobicity and phylogeny. *Nature* **373**:565–566.
- PAGEL, M. D., and P. H. HARVEY. 1988. Recent developments in the analysis of comparative data. *Q. Rev. Biol.* **63**:413–440.
- PORTER, T. D. 1995. Correlation between codon usage, regional genomic nucleotide composition, and amino acid composition in the cytochrome P-450 gene superfamily. *Biochim. Biophys. Acta* **1261**:394–400.
- RODRIGUEZ-TRELLES, F., R. TARRIO, and F. J. AYALA. 1999. Switch in codon bias and increased rates of amino acid substitution in the *Drosophila saltans* species group. *Genetics* **153**:339–350.
- SUEOKA, N. 1961. Compositional correlation between deoxyribonucleic acid and protein. *Cold Spring Harb. Symp. Quant. Biol.* **26**:35–43.

- . 1988. Directional mutation pressure and neutral molecular evolution. *Proc. Natl. Acad. Sci. USA* **85**:2653–2657.
- WILQUET, V., and M. VAN DE CASTEELE. 1999. The role of the codon first letter in the relationship between genomic GC content and protein amino acid composition. *Res. Microbiol.* **150**:21–32.
- WOOD, T. C., and W. R. PEARSON. 1999. Evolution of protein sequences and structures. *J. Mol. Biol.* **291**:977–995.

DAVID IRWIN, reviewing editor

Accepted June 30, 2000