J Mol Evol (2003) 56:362–370
DOI: 10.1007/s00239-002-2406-1

JOURNAL OF MOLECULAR EVOLUTION

Effects of GC Content and Mutational Pressure on the Lengths of Exons and Coding Sequences

Xuhua Xia,[1,2] Zheng Xie,[2] Wen-Hsiung Li[3]

[1] Department of Biology, University of Ottawa, Ottawa, Ontario, Canada KIN 6N5
[2] Institute of Environmental Protection, Hunan University, Changsha, China
[3] Department of Ecology and Evolution, University of Chicago, Chicago IL 60637-1573, USA

Received: 10 May 2002/Accepted: 11 October 2002

Abstract. It has been hypothesized that the length of an exon tends to increase with the GC content because stop codons are AT-rich and should occur less frequently in GC-rich exons. This prediction assumes that mutation pressure plays a significant role in the occurrence and distribution of stop codons. However, the prediction is applicable not to all exons, but only to the last coding exon of a gene and to single-exon CDS sequences. We classified exons in multiexon genes in eight eukaryotic species into three groups—the first exon, the internal, and the last exon—and computed the Spearman correlation between the exon length and the percentage GC (%GC) for each of the three groups. In only five of the species studied is the correlation for the last coding exon greater than that for the first or internal exons. For the single-exon CDS sequences, the correlation between CDS length and %GC is mostly negative. Thus, eukaryotic genomes do not support the predicted relationship between exon length and %GC. In prokaryotic genomes, CDS length and %GC are positively correlated in each of the 68 completely sequenced prokaryotic genomes in GenBank with genomic GC contents varying from 25 to 68%, except for the wall-less *Mycoplasma genitalium* and the syphilis pathogen *Treponema pallidum*. Moreover, the average CDS length and the genomic GC content are also positively correlated. After correcting for genome size, the partial correlation between the average CDS length and the genomic GC content is 0.3217 ($p < 0.025$).

Key words: GC content — Exon length — Coding sequences — Mutation pressure — Nonsense

Correspondence to: Wen-Hsiung Li; *email:* whli@uchicago.edu

Introduction

Oliver and Marin (1996) predicted that exon lengths should increase with the GC content of exons for the following reason. AT-rich exons may be the result of long-term AT-biased mutations, which would increase the chance of a stop codon such as TAA appearing in the middle of the exon and consequently reduce the exon length. On the other hand, GC-rich exons may be the result of long-term GC-biased mutations. In the extreme case when the sequence is made entirely of GC pairs, the exon would be infinitely long. Thus, assuming that random mutation is the dominant force in the evolution of protein-coding genes, we should have two expectations. First, if different segments of the genome are subject to different mutational pressure, some AT-biased and some GC-biased, then the length and the GC content of an exon should be positively correlated within each genome. Second, we should expect genomes with a low GC content to have shorter exons or CDSs than those with a high GC content.

Oliver and Marin (1996) compiled exon data from five vertebrate species and four prokaryotic species to test the first prediction. Four of the five vertebrate species and three of the four prokaryotic species are consistent with the first prediction. Oliver and Marin concluded that the predicted positive relationship between the length and the GC content of the coding sequence is supported and, consequently, that the difference in nonsense mutation rate between GC-rich and GC-poor regions has a significant effect on the length of coding sequences.

There are two problems with the study by Oliver and Marin (1996), aside from the failure to test the second prediction. First, the hypothesized relationship is applicable only to single-exon genes and the last coding exon in multiexon genes. Note that the last exon may contain a 3-untranslated region (UTR), and only the coding part of the last exon is relevant to the hypothesis. To be more explicit, suppose that an exon is the second in a three-exon gene. Also, suppose that the exon is experiencing strong AT-biased mutation pressure. According to the hypothesis, this exon will tend to have a nonsense mutation that has two consequences. One is that the exon will become the last coding exon, and the other is that the exon is split into the coding part and the 3-UTR. The length of the exon may still be the same, but the length of the coding part is shortened due to the nonsense mutation. Thus, it is important to use only the coding part of the last coding exon for testing the effect of nonsense mutation on the exon length in eukaryotic genomes.

In a slightly different scenario, suppose that an exon is the last in a three-exon gene, and that the exon is under strong GC-biased mutation pressure that tends to mutate the stop codon to a sense codon. Such a mutation will extend the length of the coding part of the exon downstream until another termination codon is encountered, but the exon will still be the last coding exon of the three-exon gene. It is also important to remember that, if the sequence downstream of the original stop codon is AT-rich, then a new stop codon may soon be found downstream and the exon will be lengthened only slightly. In contrast, if the sequence downstream is GC-rich, then the exon may be lengthened greatly simply because it is rare to encounter a stop codon in a GC-rich sequence.

The above two mutation scenarios will contribute to a positive correlation between the length and the GC content of the exon only when the exon is the last exon or in a single-exon gene and when its length is defined specifically as the length of the coding part of the exon. In contrast, an internal exon of a multiexon gene apparently has never accepted a nonsense mutation and therefore will not follow the prediction. If the length and the GC content of internal exons are positively correlated, it is because of other factors, not because of nonsense mutations mediated by the variation in GC content. For this reason, the data analysis by Oliver and Marin (1996), who lumped all the exons in their analysis and do not seem to have excluded the untranslated regions, is not relevant to testing the hypothesis that the different frequencies of nonsense mutations in GC-rich and GC-poor regions affects the length of coding sequences.

The second problem with the previous study concerns the frequency of codon substitutions involving stop codons. It is known that codon mutations involving very different amino acids are subject to strong purifying selection and consequently contribute little to the evolution and divergence of protein-coding genes (Kimura 1983; Xia 1998; Xia and Li 1998). For example, based on a comparative study of 58 potentially orthologous protein-coding genes from human, mouse, and rat, codon substitutions involving a large Grantham (1974) distance seldom occur, i.e., they contribute little to the evolution of protein-coding genes. Although there has been no quantitative measure of the effect of a nonsense mutation, our intuition is that the effect should most likely be more drastic than the most radical amino acid replacement. In other words, nonsense mutations are expected to have major functional effects and consequently should contribute little to the evolution of protein-coding genes. Therefore, we should not expect Oliver and Marin's prediction to be strongly supported by data from nuclear genes of eukaryotic genomes, unless the frequency of nonsense mutations is substantially higher than what we generally believe. However, the prediction may hold true for prokaryotic genomes, in which the mutation rate (and consequentially the substitution rate) is typically much higher than in eukaryotic genomes. In short, the predicted correlation between the length and the GC content of exons should be stronger in prokaryotic genomes than in eukaryotic genomes.

In this paper, we performed two critical tests of the hypothesis of Oliver and Marin (1996). First, we compiled multiexon genes from a variety of eukaryotic species, classified the exons into the first, the internal, and the last coding exons, evaluated the relationship between the length and the GC content of exons for each of the three exon classes, and tested whether the relationship is consistently stronger in the last coding exon than in the first and internal exons. We also compiled single-exon CDS sequences and tested whether the relationship between the CDS length and the GC content is stronger in the single-exon CDS sequences than in the first or middle exons. Second, we compiled the average CDS length and genomic GC content from 68 complete prokaryotic genomes and tested whether genomes with a low GC

content have shorter CDSs than genomes with a high GC content.

Materials and Methods

This study is based on two sets of data. One set is from 10 eukaryotic species: human (*Homo sapiens*), rat (*Rattus norvegicus*), mouse (*Mus musculus*), cow (*Bos taurus*), chicken (*Gallus gallus*), frog (*Xenopus laevis*), fruit fly (*Drosophila melanogaster*), nematode (*Caenorhabditis elegans*), yeast (*Saccharomyces cerevisiae*), and a plant (*Arabidopsis thaliana*). The other set is from 68 complete prokaryotic genomes with genomic GC contents varying from 25 to 68%.

Exon Length and Percentage GC (%GC) in Multiexon Genes of Eukaryotic Species

We need complete exons to evaluate the effect of GC content on exon length. We retrieved genomic sequences for human (chromosome 1), fruit fly (first scaffold; AE002566), nematode (chromosome 1), yeast (all 14 chromosomes), and *Arabidopsis thaliana* (chromosome 1) from the NCBI server at http://www.ncbi.nlm.nih.gov/PMGifs/Genomes/euk_g.html and then classed all CDS sequences into first exon, internal exon, last exon, and single-exon CDS sequences using DAMBE (Xia 2000; Xia and Xie 2001). The reason for choosing the first chromosome or the first scaffold is simply for convenience. For species without genomic sequences available, we first extracted all completely sequenced nuclear CDSs from the ACNUC server at http://pbil.univ-lyon1.fr/search/query.html, with the specification of the species name together with type = CDS AND NOT keyword = partial AND NOT organelle = mitochondrion. The resulting file in GenBank format was then parsed by DAMBE to produce the first exon, internal exon, last exon, and single-exon CDS sequences. The predicted effect of nonsense mutations in GC-rich and GC-poor regions on the exon length should be applicable only to the last exon and single-exon CDS sequences.

Complete Prokaryotic Genomes

Sixty-eight complete prokaryotic genomes were retrieved from the genome database at http://www.ncbi.nlm.nih.gov/PMGifs/Genomes/bact.html. There are three advantages of using complete genomes. First, it will avoid sampling bias. For example, if GC-rich DNA is more difficult to sequence than GC-poor DNA, then researchers may avoid sequencing long GC-rich CDSs. Consequently, the long GC-rich CDSs may be less represented in GenBank and this would bias the estimation of the relationship between the length and the GC content of the coding sequences. Second, retrieving all CDSs from one particular bacterial species may occasionally include plasmid CDSs, which should not be lumped with bacterial genomic CDSs in estimating the relationship between the length and the GC content of CDSs. Using complete genomes avoids this problem of including plasmid CDSs. Third, with complete genomes, we can test whether genomes with a high genomic GC content tend to have longer CDSs than those with a low genomic GC content.

Prokaryotic protein genes are typically made up of a single exon. However, some genes may occasionally be divided due to the insertion of a prophage (e.g., the *wrbA* gene in GenBank LOCUS AP000422) or other sequences (e.g., the *Stx2* variant B subunit in GenBank LOCUS AB017524). The *prfB* gene in GenBank LOCUS AE000372 represents another kind of deviation. It has an in-frame premature UGA termination codon located within the sequence, and a naturally occurring +1 frameshift is required for its synthesis. In a sense, the gene CDS is made up of two segments separated by a single nucleotide. All prokaryotic CDSs divided in any way are considered exceptions and simply discarded. In short, only strict single-exon genes in prokaryotic genes are used.

Statistical Methods

Under the random distribution of nucleotides, the length of a coding sequence is expected to increase with its GC content (Oliver and Marin 1996). However, the relationship is not expected to be linear, for the following reason. The frequency of the 64 codons, if assembled randomly, is expected to be $P_i \cdot P_j \cdot P_k$, with P_i, P_j, and P_k being the frequencies of nucleotides i, j, and k ($i, j, k \in$ {A, C, G, T, or U}) in the RNA or DNA sequences. This is perhaps true before the fixation of the genetic code in evolutionary history. The distribution of the peptide length (L) follows the geometric distribution:

$$ppp \tag{1}$$

where p is the proportion of stop codons and is equal to 3/64 for the universal genetic code with the assumption of equal nucleotide frequencies and equal codon usage. The expected mean and variance of the exon length are then

$$\begin{aligned} E(L) &= \frac{1}{p} \\ \mathrm{Var}(L) &= \frac{(1-p)}{p^2} \end{aligned} \tag{2}$$

If we designate the four nucleotide frequencies P_A, P_C, P_G, and P_T, and assume that $P_A = P_T$ and $P_C = P_G$, then the relationship between L and P_{CG} ($= P_C + P_G$) is

$$xxxx \tag{3}$$

It is obvious that the relationship between L and P_{GC} is not linear, and the use of the Pearson correlation coefficient is inappropriate. For this reason, we compute the nonparametric Spearman's rank correlation (R_s) between the length and the %GC of the coding sequence.

For the 68 complete prokaryotic genomes, we evaluated not only the relationship between the length and the %GC of CDSs within each genome, but also the relationship between the genomic %GC and the average CDS length. To avoid phylogenetic dependence, we used the independent contrasts (Felsenstein 1985) implemented in the CONTRAST program in an earlier version of PHYLIP (Felsenstein 1993) and the variance-partitioning method (Lynch 1991) implemented in the CONTRAST program in the most recent version of PHYLIP. The 16S rRNA gene from each of the genomes was extracted, aligned, and used to build the phylogenetic tree for the two comparative methods. The neighbor-joining method (Saitou and Nei 1987) with the paralinear distance (Lake 1994) was used to build the phylogenetic tree, which was checked to make sure that no branch length is negative. We used the paralinear distance because the sequences differ greatly in nucleotide frequencies, and the paralinear distance can accommodate such nonstationary features of the substitution process.

Results and Discussion

Multiexon and Single-Exon Genes in Eukaryotic Species

All Spearman correlation coefficients (R_s) between the length and the %GC of exons are positive, and highly significant in most cases, for multiexon genes

Table 1. Relationship between exon length and %GC, measured by the Spearman rank correlation, for three classes of exons: the first exon (first), internal exon (internal), and last exon (last)[a]

Species	Class	N_{Exon}	$L \pm$ SD	%GC $\pm$ SD	R_S	p
H. sapiens (chromosome 1)	First	3,015	162.0 ± 255.9	54.8 ± 11.6	0.2773	0.0000
	Internal	20,334	142.9 ± 152.8	50.3 ± 9.0	0.1518	0.0000
	Last	3,007	196.9 ± 248.4	50.1 ± 10.4	0.2735	0.0000
R. norvegicus	First	465	205.2 ± 409.1	56.2 ± 8.7	0.1420	0.0021
	Internal	2,871	127.1 ± 96.1	51.8 ± 7.0	0.1766	0.0000
	Last	495	278.1 ± 522.8	51.7 ± 8.7	0.3083	0.0000
M. musculus	First	2,161	194.9 ± 363.4	57.1 ± 9.4	0.2407	0.0000
	Internal	12,567	145.1 ± 171.6	53.1 ± 7.4	0.1304	0.0000
	Last	2,163	246.1 ± 353.3	52.0 ± 9.0	0.3151	0.0000
B. taurus	First	162	202.7 ± 301.8	58.5 ± 10.2	0.2073	0.0081
	Internal	544	132.8 ± 70.0	54.2 ± 8.4	0.2149	0.0000
	Last	167	218.5 ± 243.8	52.8 ± 11.2	0.3403	0.0000
G. gallus	First	231	226.8 ± 442.8	60.6 ± 11.0	0.0310	0.6391
	Internal	1,412	143.2 ± 122.0	52.1 ± 9.3	0.0659	0.0133
	Last	240	277.7 ± 504.6	52.7 ± 9.8	0.2681	0.0000
X. laevis	First	66	213.4 ± 351.1	49.6 ± 8.6	0.1819	0.1439
	Internal	181	152.3 ± 82.1	46.9 ± 4.9	0.1841	0.0131
	Last	66	375.7 ± 401.5	45.7 ± 6.4	0.0546	0.6633
C. elegans (chromosome 1)	First	2,125	154.3 ± 212.9	42.3 ± 7.5	0.0377	0.0820
	Internal	9,664	219.3 ± 205.0	42.1 ± 5.8	0.0329	0.0012
	Last	2,125	208.9 ± 205.6	43.0 ± 6.6	0.0253	0.2445
D. melanogaster (scaffold 1; AE002566)	First	819	337.9 ± 436.6	54.6 ± 7.2	0.3164	0.0000
	Internal	1,897	438.5 ± 585.0	55.2 ± 5.1	0.2043	0.0000
	Last	811	495.1 ± 610.0	55.2 ± 5.9	0.2720	0.0000
S. cerevisiae (genome)	First	245	300.8 ± 480.6	39.2 ± 8.4	0.0441	0.4924
	Internal	7	157 ± 158.5	39.1 ± 4.8	−0.0360	0.9389
	Last	280	1,231.2 ± 1,399.9	40.7 ± 4.0	−0.2579	0.0000
A. thaliana (chromosome 1)	First	5,387	320.3 ± 371.8	45.5 ± 5.4	0.0370	0.0066
	Internal	24,051	167.6 ± 195.7	43.2 ± 4.3	0.0180	0.0053
	Last	5,387	328.3 ± 354.5	44.2 ± 4.8	0.0663	0.0000

[a] N_{Exon}—number of exons in the class; L—mean exon length; P_{GC}—proportion of G + C; SD—standard deviation (it is not meaningful to compare variability using the standard error with different sample sizes); R_s—Spearman rank correlation; p—probability of mistakenly rejecting the null hypothesis of $R_s = 0$.

(Table 1). This is consistent with the previous finding (Oliver and Marin 1996). However, this universally positive correlation does not support the hypothesized mechanism (Oliver and Marin 1996) responsible for the correlation. According to the hypothesis, the positive correlation is due to AT-biased mutations leading to nonsense mutations and shortened last coding exons or GC-biased mutations leading to the loss of stop codons and consequent lengthening of the last coding exons. As we argued above, this hypothesis is applicable only to the last coding exon in multiexon genes or to single-exon CDS sequences. It predicts that the positive correlation between the length and the GC content of an exon should be stronger in the last coding exon, or in single-exon CDS sequences, than in the first and internal exons if nonsense mutations play a significant role in determining exon length.

The R_s value for the last coding exon is not consistently larger than those for the other two classes of exons (Table 1). Of the 10 eukaryotic species, the 5 warm-blooded mammalian and avian species are consistent in the predicted direction (Table 1). We have also compiled similar, but less complete, data for 10 other mammalian and avian species and the results are consistent. We may conclude that the prediction is generally true for mammalian and avian species.

The prediction, however, fails badly with the five other species in Table 1. As the prediction should hold not just for mammalian and avian species, but for all species, we conclude that the prediction is not universally supported. As the number of exons in each exon class is quite large, the lack of a clear-cut result suggests that the effect of nonsense mutations on exon length is weak in eukaryotic species.

One might argue that avian and mammalian species, having higher body temperatures than other species, might have higher mutation rates and consequently would be better candidates for supporting the prediction. This revised hypothesis unfortunately is also not true. We have compiled similar data for the zebrafish (*Danio rerio*) and several salmon species (*Oncorhynchus* sp.) and the result is similar to that for warm-blooded mammals and birds, although the revised hypothesis predicts that the cold-blooded fish

Table 2. Relationship between exon length and %GC, measured by the Spearman rank correlation, for single-exon CDS sequences[a]

Species	N_{Exon}	$L \pm SD$	%GC ± SD	R_S	P
H. sapiens (chromosome 1)	417	587.3 ± 458.8	52.1 ± 10.2	0.2796	0.0000
R. norvegicus	6,403	1,541.3 ± 1,304.8	52.7 ± 5.9	0.0074	0.5557
M. musculus	15,114	1,410.0 ± 1,222.7	52.6 ± 6.2	0.0136	0.0938
B. taurus	1,635	1,390.2 ± 1,237.3	53.3 ± 7.5	−0.0687	0.006
G. gallus	1,928	1,501.9 ± 1,380.5	53.5 ± 8.5	−0.2222	0.0000
X. laevis	1,987	1,432.4 ± 1,063.3	48.0 ± 4.8	−0.1833	0.0000
C. elegans (chromosome 1)	65	651.5 ± 562.1	46.0 ± 6.5	−0.0428	0.7352
D. melanogaster (scaffold 1)	306	814.4 ± 601.2	55.4 ± 5.0	0.0162	0.7772
S. cerevisiae (genome)	6,025	1,409.1 ± 1,097.3	40.2 ± 3.8	−0.1858	0.0000
A. thaliana (chromosome 1)	1,370	1,042.7 ± 739.3	45.2 ± 4.3	−0.2010	0.0000

[a] For abbreviations see Table 1, footnote a.

species should be different from the warm-blooded mammalian and avian species and similar to the last five species in Table 1. Thus, the revised hypothesis has to be rejected unless one wants to make an unsubstantiated argument that the fish species also have higher mutation rates.

It is not clear why there should be a universal positive correlation between the length and the %GC of exons that are not the last coding exon. One may argue that the last coding exon may also become the first exon or an internal exon of a multiexon gene. If this is true, then the positive correlation between the length and the %GC of the last coding codon may be carried over to the first and internal exons, so that we will observe a positive correlation between the length and the %GC not only for the last coding exon, but also for the other two classes of exons. However, this argument appears to be far-fetched and is very difficult to substantiate empirically.

One interesting observation is that the last exons in *S. cerevisiae* are on average four times longer than the first exon and eight times longer than the internal exons, whereas little difference in exon length exists between the last exon and the first and internal exons for all other species (Table 1). It is unclear why *S. cerevisiae* should have such extraordinarily long last exons. We note that the last exons in *S. cerevisiae* do have, on average, a higher %GC than the first and internal exons, suggesting that the mutation and selection during the evolution of the last exons are in favor of increasing GC. However, if we want to make an argument that it is the GC-biased substitution that leads to the loss of stop codons and the lengthening of the last exon, then we should expect a positive correlation between the last exon length and the %GC. The correlation, however, is highly significantly negative ($r_s = -0.2579$, $p = 0.0000$; Table 1). Thus, the long last coding exons in *S. cerevisiae* need to be explained by factors other than increased GC content.

The single-exon genes do not support the prediction at all (Table 2), with more than half of the R_s values being negative, and most being smaller than the corresponding R_s values for the first and internal exons (Table 1). Even the single-exon CDS data for the warm-blooded species do not support the prediction. Only one (human) of the five warm-blooded species has a R_s value from the single-exon CDS sequences higher than those from the first and internal exons (Tables 1 and 2). We therefore conclude that the predicted effect of %GC on exon length mediated by the loss or gain of stop codons is minimal in eukaryotic species.

One point that we wish to make is that a less careful approach to data analysis could generate very strong support for the prediction. For example, if we argue that the predicted relationship between the exon length and the %GC should apply to both the last coding exon and the exons in single-exon CDS sequences and lump the two kinds of heterogeneous exons together, then we would have a strong positive correlation between the exon length and the %GC for all species except *S. cerevisiae*. This is because the exons in single-exon CDS sequences are generally much longer and have a higher %GC than the last coding exon, except for *S. cerevisiae*, in which the last coding exons are almost as long as the exons in single-exon CDSs, and the %GC is slightly higher in the former than the latter (Tables 1 and 2).

Prokaryotic Genomes

Within-Species Comparisons. With the exception of *M. genitalium* and *Treponema pallidum*, the R_s values for the prokaryotic genomes are positive and, in most cases, highly significant (Table 3). The result is consistent with the prediction from the hypothesis of Oliver and Marin (1996). The magnitude of the R_s value depends significantly ($p < 0.001$) on the genome size, with larger genomes having higher R_s values, although the relationship is not linear (Fig. 1). One possible cause for the correlation is that large genomes may have greater variation in %GC. The correlation between the log-transformed genome size and the standard deviation of %GC is indeed positive ($r = 0.192$) but it is not significant ($p > 0.05$).

Table 3. Relationship between CDS length and %GC for the 71 complete prokaryotic genomes (chromosomes), measured by the Spearman rank correlation[a]

Species	N_{CDS}	$L \pm$ SD	%GC $\pm$ SD	R_S	p
A. tumefaciens strain C58					
Cereon, circular	2785	896.4 ± 610.4	59.4 ± 3.4	0.3954	0.0000
Cereon, linear	1876	993.1 ± 614.4	59.3 ± 3.5	0.2961	0.0000
Dupont, circular	2721	933.3 ± 621.4	59.4 ± 3.3	0.3559	0.0000
Dupont, linear	1833	1034.7 ± 626.0	59.4 ± 3.3	0.2529	0.0000
Bacillus halodurans	4066	879.9 ± 571.7	43.7 ± 3.7	0.3278	0.0000
Bacillus subtilis	4104	893.4 ± 775.8	43.3 ± 4.6	0.3720	0.0000
Borrelia burgdorferi	851	1002.8 ± 675.4	28.4 ± 3.9	0.1754	0.0000
Brucella melitensis (chromosome I)	2059	884.2 ± 622.5	57.6 ± 3.7	0.3371	0.0000
Brucella melitensis (chromosome II)	1139	912.8 ± 562.9	57.7 ± 3.4	0.3003	0.0000
Buchnera sp. APS	564	988.0 ± 629.3	27.3 ± 3.7	0.0515	0.2216
Campylobacter jejuni	1654	940.1 ± 596.0	30.6 ± 3.7	0.1349	0.0000
Caulobacter crescentus	3737	937.7 ± 641.9	67.3 ± 3.3	0.2253	0.0000
Chlamydia muridarum	909	1067.2 ± 888.1	40.6 ± 2.6	0.1042	0.0017
Chlamydia pneumoniae AR39	1112	985.3 ± 731.4	40.5 ± 3.4	0.3549	0.0000
Chlamydia trachomatis	895	1049.7 ± 733.1	41.5 ± 2.3	0.1166	0.0005
Chlamydophila pneumoniae CWL029	1054	1033.8 ± 717.3	40.9 ± 2.8	0.3093	0.0000
Chlamydophila pneumoniae J138	1069	1031.0 ± 714.2	40.9 ± 2.8	0.3110	0.0000
Clostridium acetobutylicum ATCC824	3672	921.0 ± 662.6	31.1 ± 3.3	0.2077	0.0000
Deinococcus radiodurans R1					
Chromosome 1	2629	909.1 ± 577.1	67.2 ± 4.3	0.1865	0.0000
Chromosome 2	368	1046.0 ± 613.1	67.0 ± 5.1	0.0728	0.1635
Escherichia coil K12	4279	957.2 ± 624.9	51.0 ± 4.9	0.3248	0.0000
Escherichia coli O157H7	5361	903.5 ± 691.7	50.6 ± 5.6	0.3174	0.0000
Escherichia coli O157H7 EDL933	5324	909.8 ± 695.5	50.4 ± 5.9	0.3248	0.0000
Haemophilus influenzae Rd	1714	941.1 ± 617.3	38.4 ± 3.5	0.2041	0.0000
Helicobacter pylori 26695	1576	957.5 ± 718.8	39.1 ± 3.8	0.2043	0.0000
Helicobacter pylori J99	1491	997.1 ± 742.6	39.6 ± 3.5	0.1865	0.0000
Lactococcus lactis subsp. *lactis*	2267	883.6 ± 616.1	35.7 ± 3.5	0.2382	0.0000
Listeria innocua Clip11262	2980	900.0 ± 638.8	37.4 ± 3.3	0.2254	0.0000
Listeria monocytogenes strain EGD	2855	920.3 ± 630.0	38.0 ± 3.4	0.2449	0.0000
Mesorhizobium loti	6746	907.3 ± 639.5	63.0 ± 3.5	0.2794	0.0000
Mycobacterium leprae strain TN	2720	922.2 ± 693.0	58.4 ± 3.5	0.2263	0.0000
Mycobacterium tuberculosis CDC1551	4187	952.9 ± 765.5	65.2 ± 3.5	0.2366	0.0000
Mycobacterium tuberculosis H37Rv	3911	1022.0 ± 791.7	65.5 ± 3.4	0.1859	0.0000
Mycoplasma genitalium	484	1093.5 ± 790.1	31.5 ± 3.7	−0.0300	0.5099
Mycoplasma pneumoniae	689	1046.9 ± 750.6	40.3 ± 4.9	0.1767	0.0000
Mycoplasma pulmonis	776	1117.0 ± 915.2	26.9 ± 3.8	0.1194	0.0009
N. meningitidis serogroup A strain Z2491	2121	854.0 ± 635.3	51.7 ± 7.1	0.3650	0.0000
N. meningitidis serogroup B strain MC58	2067	870.2 ± 697.9	51.4 ± 7.6	0.3951	0.0000
Nostoc sp. PCC 7120	5366	982.5 ± 815.3	42.0 ± 4.1	0.1898	0.0000
Pasteurella multocida	2015	997.7 ± 695.1	40.7 ± 3.4	0.2554	0.0000
Pseudomonas aeruginosa	5567	1005.3 ± 748.4	66.7 ± 4.0	0.1979	0.0000
Ralstonia solanacearum	3435	950.6 ± 738.3	67.0 ± 4.6	0.2427	0.0000
Rickettsia conorii Malish 7	1374	746.4 ± 679.3	32.0 ± 4.1	0.2460	0.0000
Rickettsia prowazekii strain Madrid E	835	1006.3 ± 694.5	30.0 ± 3.4	0.1909	0.0000
S. enterica subsp. *enterica* serovar Typhi	4600	916.1 ± 650.3	52.1 ± 5.6	0.3511	0.0000
S. typhimurium LT2	4451	947.6 ± 668.3	52.4 ± 5.4	0.3320	0.0000
Sinorhizobium meliloti 1021	3332	942.3 ± 630.6	63.0 ± 3.0	0.3008	0.0000
Staphylococcus aureus strain Mu50	2714	889.1 ± 752.3	33.0 ± 3.3	0.2624	0.0000
Staphylococcus aureus subsp. *aureus* N315	2594	906.9 ± 762.2	33.0 ± 3.3	0.2740	0.0000
Streptococcus pneumoniae R6	2043	868.2 ± 668.0	39.9 ± 4.5	0.2973	0.0000
Streptococcus pneumoniae TIGR4	2094	853.5 ± 719.7	39.5 ± 5.0	0.3361	0.0000
Streptococcus pyogenes	1697	914.7 ± 619.9	38.6 ± 3.7	0.2524	0.0000
Synechocystis sp. PCC 6803	3167	982.0 ± 766.8	48.2 ± 5.0	0.1691	0.0000
Thermotoga maritime	1858	947.5 ± 590.2	46.1 ± 3.4	0.2167	0.0000
Treponema pallidum	1036	1020.9 ± 669.2	52.9 ± 4.3	−0.0798	0.0102
Ureaplasma urealyticum	613	1119.8 ± 1122.8	25.8 ± 3.8	0.0056	0.8890
Vibrio cholerae					
Chromosome I	2742	949.6 ± 711.9	47.4 ± 4.3	0.3719	0.0000
Chromosome II	1093	832.4 ± 672.6	46.0 ± 4.8	0.5151	0.0000
Xylella fastidiosa	2766	808.2 ± 759.2	52.2 ± 6.2	0.3607	0.0000

Table 3. Continued

Species	N_{CDS}	$L \pm$ SD	%GC $\pm$ SD	R_S	p
Yersinia pestis	3994	974.9 ± 734.6	47.9 ± 4.9	0.3412	0.0000
Methanothermobacter thermautotrophicus	1871	845.8 ± 584.1	49.9 ± 4.4	0.3498	0.0000
Sulfolobus solfataricus	2977	850.0 ± 514.0	36.4 ± 4.5	0.0979	0.0000
Aeropyrum pernix	1840	844.0 ± 564.7	56.9 ± 5.7	0.0904	0.0001
Archaeoglobus fulgidus	2416	833.5 ± 551.7	48.7 ± 4.0	0.3345	0.0000
Halobacterium sp. NRC-1	2075	862.2 ± 561.9	67.9 ± 5.2	0.2513	0.0000
Pyrococcus abyssi	1767	912.9 ± 562.5	45.0 ± 3.5	0.1148	0.0000
Pyrococcus horikoshii	1798	891.5 ± 616.8	42.1 ± 3.5	0.1341	0.0000
Sulfolobus tokodaii	2826	808.2 ± 535.3	33.4 ± 4.4	0.1298	0.0000
Thermoplasma volcanium	1499	903.8 ± 596.4	40.9 ± 3.4	0.1949	0.0000
Methanococcius jannaschii	1729	851.2 ± 601.9	31.7 ± 3.8	0.1350	0.0000
Thermoplasma acidophilum	1482	922.5 ± 589.0	46.7 ± 4.0	0.3077	0.0000

[a] For abbreviation see Table 1, footnote a. The last 11 species belong to Archaea and the rest to Bacteria.

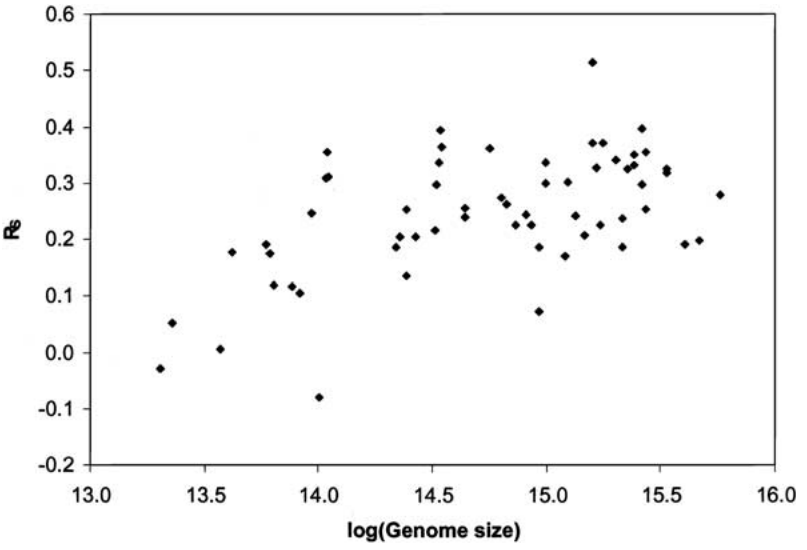

Fig. 1. Large genomes tend to have a higher correlation, measured by Spearman rank correlation (R_S), between CDS length and %GC than small genomes.

The genomic %GC is also positively and highly significantly ($p < 0.001$) correlated with the genome size (Fig. 2). This offers another explanation for the observed positive correlation between R_s and genome size. According to Eq. (3), the effect of P_{GC} variation on L increases with L because L increases with P_{GC} at an increasing rate. For example, a change in the %GC from 20 to 30% will cause little change in L, but a change from 60 to 70% will change L dramatically. In other words, the same within-genome variation in %GC is expected to affect the CDS length in genomes with a large %GC more than in genomes with a small %GC.

It is not clear why large genomes have a higher GC content than small genomes do in prokaryotes. One possible explanation involves integrating DNA from external sources in the environment such as in lateral gene transfer (Boucher and Doolittle 2000). If larger genomes result from integrating external DNA, and if DNA fragments of high GC content persist for a longer period than DNA fragments of low GC content, then DNA fragments of high GC content in the environment would have a better chance of being taken up by a prokaryotic cell. Therefore, prokaryotes integrating many GC-rich fragments from the environment would have a large genome size and a high GC content. This is of course just a speculation

Between-Species Comparisons. If a high GC content favors the loss of stop codons and the lengthening of coding sequences according to the hypothesis of Oliver and Marin (1996), then genomes with a high genomic %GC are expected to have longer CDS sequences than genomes with a low %GC. We use genomic CDS length and genomic %GC below to designate the mean of all CDS lengths and the mean of all CDS %GC, respectively, for each genome. Only the 57 eubacterial genomes were used for assessing the relationship between genomic CDS length and genomic %GC.

Genomic CDS length and genomic %GC are negatively, although not significantly, correlated in the 57 eubacterial genomes (Fig. 3), contrary to the

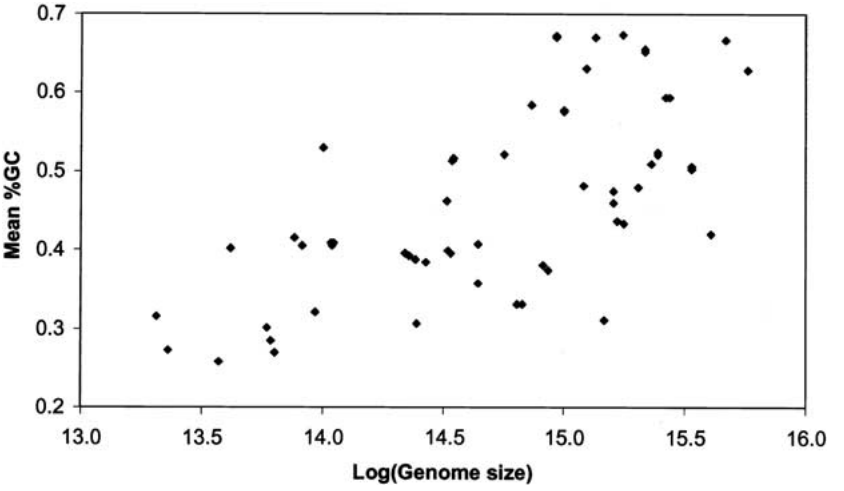

Fig. 2. GC content increases with genome size in prokaryotic genomes.

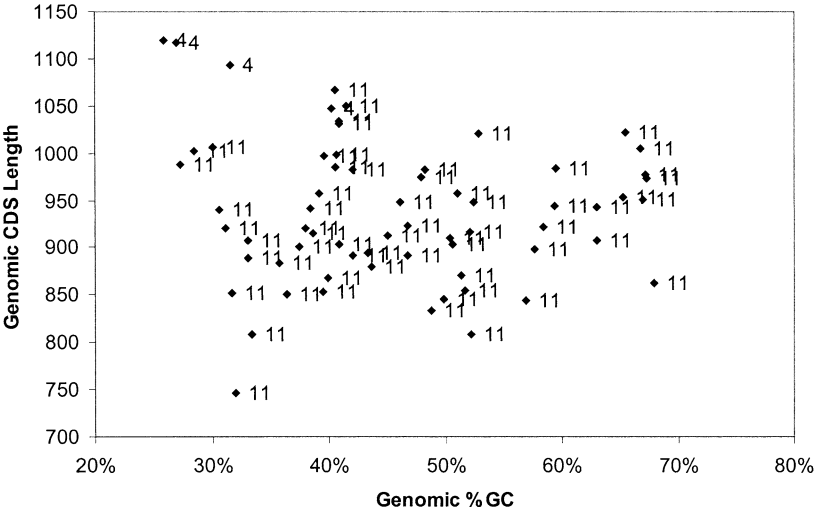

Fig. 3. Relationship between genomic CDS length and genomic %GC. Numbers (4 and 11) indicate the translation table for the two genetic codes.

prediction of a positive correlation. This result is paradoxical and does not seem to be compatible with the results in Table 3. However, there are two easily identifiable factors confounding the test of the prediction. First, the data included four Mycoplasma species (i.e., *Mycoplasma genitalium*, *M. pneunoniae*, *M. pulmonis*, and *Ureaplasma urealyticum*) that use a genetic code (trans_table = 4 with only two stop codons, UAA and UAG) different from that for the rest of the eubacterial species (trans_table = 11 with three stop codons, UAA, UAG, and UGA). The probability of encountering a stop codon is the summation of the probability of encountering UAA and UAG for the Mycoplasma species, whereas it is the summation of the probability of encountering UAA, UAG, and UGA for the rest of the species. Clearly, given the same genomic %GC, the Mycoplasma species should have a lower probability of encountering a stop codon, and consequently should have a longer genomic CDS length, than the rest of the bacterial species. It is therefore not surprising to see that the four Mycoplasma species (labeled 4 in Fig. 3) have longer genomic CDSs than the rest of the eubacterial species. The correlation between the genomic CDS length and the genomic %GC becomes positive when the four Mycoplasma species are excluded from the data set.

The second confounding factor is the correlation of genome size with both genomic CDS length and genomic %GC. Genome size is positively correlated with genomic %GC (Fig. 2) but negatively correlated with genomic CDS length (−0.346; $p = 0.018$). Thus, it is necessary to control for the effect of genome size when measuring the relationship between genomic CDS length and genomic %GC. This can be done by calculating the partial correlation between CDS length and %GC, while holding genome size constant. From the independent contrasts based on the topology of 48 eubacterial species (Fig. 4), we obtain the correlation coefficients between CDS length and %GC ($r_{\mathrm{CDSLen,GC}}$), between CDS length and genome size ($r_{\mathrm{CDSLen,GenSize}}$), and between %GC and genome

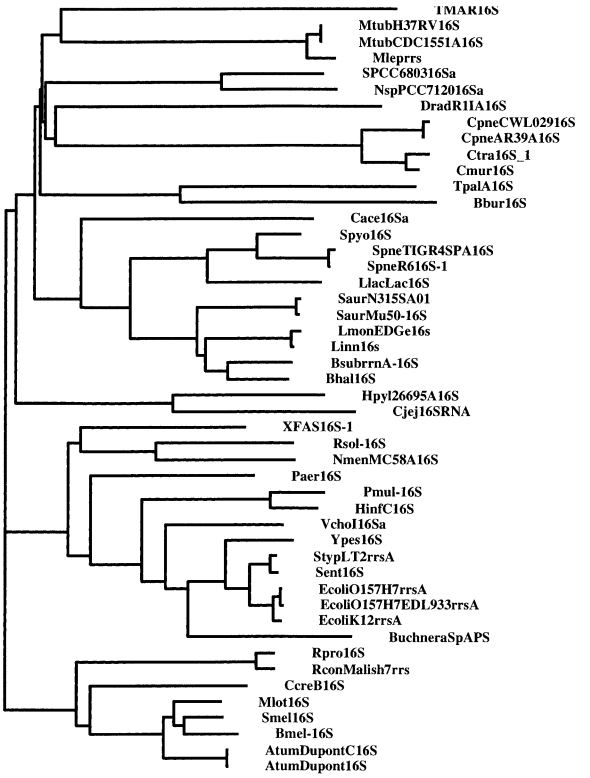

Fig. 4. Phylogenetic tree of 48 eubacterial species (after excluding the four Mycoplasma genomes and genomes that do not have annotated 16S rRNA). Based on the 16S rRNA gene sequences and the neighbor-joining method with the paralinear distance. The first letter for each OTU is the initial of the generic name, and the next three letters are the first three letters of the species name.

size ($r_{GC,GenSize}$) as −0.1266, −0.1572, and 0.6787, respectively. The partial correlation between CDS length and %GC while controlling for the effect of genome size is 0.3217, with a p value between 0.01 and 0.025 (one-tailed test). Alternatively, one might assume the lack of a phylogenetic component in determining the relationship among the three variables and just regress the CDS length on the log-transformed genome size and the %GC. The resulting slope is 194.82 ($p = 0.0456$) for the %GC and −45.38 ($p = 0.01823$) for the log-transformed genome size.

Our results show a clear difference between eukaryotic (Tables 1 and 2) and prokaryotic (Table 3) genomes in the effect of GC content on exon (CDS) length. This is not difficult to understand given the differential mutation rate (and substitution rate) between eukaryotic and prokaryotic genomes. We mentioned in the Introduction that nonsynonymous substitutions of a large effect (i.e., involving a large Grantham distance) contribute little to the evolution of protein-coding genes in three mammalian genomes (human, mouse, and rat). This might also be true for other eukaryotic genomes. Based on this observation, we have suggested that a nonsense mutation is likely to have a much more drastic effect on protein function than the most radical amino acid replacement and, consequently, would contribute even less to the evolutio of protein-coding genes in eukaryotic genomes. In contrast, prokaryotic genomes typically experience a much higher mutation rate (and, consequently, substitution rate) than eukaryotic genomes. This led us to suggest that the predicted correlation between the length and the GC content of exons (CDSs), mediated by nonsense mutations, should be stronger in prokaryotic genomes than in eukaryotic genomes. Our results are consistent with the hypothesized difference between prokaryotic and eukaryotic genomes.

In summary, the prediction by Oliver and Marin (1996) is largely consistent with the prokaryotic genomes but not with the eukaryotic genomes. However, nature clearly does not allow a GC-rich prokaryote to have consistently long CDSs or GC-poor prokaryotes to have consistently short CDSs. Instead, all prokaryotic genomes, regardless of genomic GC content, have a set of CDSs with roughly the same length distribution.

Acknowledgments. This study was supported by a grant fromt the University of Ottawa to X.Y. Chinese Ministry of Education Grant 99K68027 to Z.X. and NIH Grant GM30998 to W.H.L. We are grateful to A. Nekrutenko and Z. Gu for discussion and references and H. Kong for assistance.

References

Boucher Y, Doolittle WF (2000) The role of lateral gene transfer in the evolution of isoprenoid biosynthesis pathways. Mol Microbiol 37:703–716

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

Felsenstein J (1993) PHYLIP 3.5 (phylogeny inference package). Department of Genetics, University of Washington, Seattle

Grantham R (1974) Amino acid difference formula to help explain protein evolution. Science 185:862–864

Kimura M (1983) The neutral theory of molecular evolution. Cambridge University Press, Cambridge

Lake JA (1994) Reconstructing evolutionary trees from DNA and protein sequences: Paralinear distances. Proc Nat Acad Sci USA 91:1455–1459

Lynch M (1991) Methods for the analysis of comparative data in evolutionary biology. Evolution 45:1065–1080

Oliver JL, Marin A (1996) A relationship between GC content and coding-sequence length. J Mol Evol 43:216–223

Saitou N, Nei aM (1987) The neighbor-joining method: A new method for reconstructing phylogenetic trees. Mol Biol Evol 4:406–425

Xia X (1998) The rate heterogeneity of nonsynonymous substitutions in mammalian mitochondrial genes. Mol Biol Evol 15:336–344

Xia X (2000) Data analysis in molecular biology and evolution. Kluwer Academic, Boston

Xia X, Li W-H (1998) What amino acid properties affect protein evolution? J Mol Evol 47:557–564

Xia X, Xie Z (2001) DAMBE: Data analysis in molecular biology and evolution. J Hered 92:371–373