J Mol Evol (1996) 43:216–223

JOURNAL OF MOLECULAR EVOLUTION

A Relationship Between GC Content and Coding-Sequence Length

José L. Oliver,[1] Antonio Marín[2]

[1] Departamento de Genética, Instituto de Biotecnología, Facultad de Ciencias, Universidad de Granada, E-18071-Granada, Spain
[2] Departamento de Genética, Facultad de Biología, Universidad de Sevilla, Aptdo. 1095, E-41080-Sevilla, Spain

Received: 15 December 1995 / Accepted: 11 March 1996

Abstract. Since base composition of translational stop codons (TAG, TAA, and TGA) is biased toward a low G+C content, a differential density for these termination signals is expected in random DNA sequences of different base compositions. The expected length of reading frames (DNA segments of sense codons flanked by in-phase stop codons) in random sequences is thus a function of GC content. The analysis of DNA sequences from several genome databases stratified according to GC content reveals that the longest coding sequences—exons in vertebrates and genes in prokaryotes—are GC-rich, while the shortest ones are GC-poor. Exon lengthening in GC-rich vertebrate regions does not result, however, in longer vertebrate proteins, perhaps because of the lower number of exons in the genes located in these regions. The effects on coding-sequence lengths constitute a new evolutionary meaning for compositional variations in DNA GC content.

Key words: Base composition — Stop-codon density — Coding-sequence length — Compositional heterogeneity

Introduction

The lengths of coding DNA segments (CDS) are known to be under both functional and structural constraints (Blake 1983, 1985; Hawkins 1988; Traut 1988). There are some indications that compositional constraints may also be involved. The concentration of genes in the GC-richest fraction of the human genome is known to be five to ten times higher than the gene density in the poorer GC regions (Bernardi 1989). The dependence on GC level of open-reading-frame (ORF) lengths and stop-codon density has been previously studied (Merino et al. 1994; Boldögkoi et al. 1995; Cebrat and Dudek 1996), and ORF occurrence has been found to be positively correlated with GC content (Guigó and Fickett 1995). However, other results are puzzling. Thus, while long genes are scarce (Duret et al. 1995), long ORFs are frequent (Guigó and Fickett 1995) in GC-rich isochores. To our knowledge, no systematic analysis has yet determined which effects could be expected from compositional constraints and which, if any, can be observed in the lengths of coding sequences.

The base composition of translational stop codons (TAA, TAG, and TGA) and of their reverse complements (TTA, CTA, and TCA) is GC-poor. In random nucleotide sequences, as the primordial ones probably were (Senapathy 1986; Naora et al. 1987; Höglund et al. 1990; White and Jacobs 1993), such compositional asymmetry leads to a differential density of stop codons according to the GC content of the sequence. In GC-poor random sequences, the stop-codon density is expected to be higher than in the GC-rich ones. Thus, the length between consecutive in-phase stop codons, and therefore the reading-frame length, is a function of sequence GC content: the higher the GC content, the lower the density of stop codons and the longer, therefore, the expected reading frames. Through the analysis of several genome databases, we show here that the variations in sequence GC content found within both eukaryotic (Bernardi et al. 1985) and prokaryotic (Nomura et al. 1987; D'Onofrio

Correspondence to: J. L. Oliver; e-mail:oliver@ugr.es

and Bernardi 1992; Sueoka 1992) genomes seem to provide a propitious environment for the segregation of short and long coding sequences in the different compositional genome regions.

Data and Methods

As far as possible, clean, nonredundant sequence databases were used. First, release 23 (July 1995) of the *Escherichia coli* Database Collection (ECDC, Wahl et al. 1994), containing 1,634 genes, was retrieved from the European Bioinformatics Institute (EBI) ftp server (ftb.ebi.ac.uk). Second, release 5 (June 1995) of the nonredundant database for *Bacillus subtilis* (NRSub, Perrière et al. 1994), containing 1,085 genes, was retrieved through the ACNUC Web Homepage (http://acnuc.univ-lyonl.fr). Third, the complete genomes of both *Haemophilus influenzae* (Fleischmann et al. 1995), containing 1,726 genes, and *Mycoplasma genitalium* (Fraser et al. 1995), containing 468 genes, were retrieved from The Institute for Genome Research Web Server (http://www.tigr.org). Lastly, we analyzed the nonredundant database of vertebrate genomic sequences described by Duret et al. (1995). This database contains entries from the human (*Homo sapiens*), cow (*Bos taurus*), mouse (*Mus musculus*), rat (*Rattus norvegicus*), and chicken (*Gallus gallus*); Duret and co-workers maintain a list with the corresponding accession numbers available through anonymous ftp from biom3.univ-lyon1.fr; the entries were retrieved from the EMBL Nucleotide Sequence Data Library (Stoehr and Cameron 1991); a total of 663 vertebrate genes, 3,728 exons, and 3,063 introns from this database were analyzed.

Sequence annotation was automatically parsed, extracting the length and the GC content of genes, exons, and introns. A few coding sequences from some of the above databases with nonstandard annotation were not included in the analysis.

Data Stratification. The samples of genes, exons, and introns in each genome were stratified in three compositional classes of approximately equal size according to GC content: GC-poor (G+C $\leqslant$ P33), GC-medium (P33 > G+C $\leqslant$ P67), and GC-rich (G+C > P67), where P33 and P67 are the 33% and 67% percentiles of the G+C distribution, respectively.

Data Analysis. Several programs from BMDP software package R.7 (BMDP Statistical Software, Inc.) were used for statistical analyses. Sample basic statistics were computed with 1D and 2D programs. The Kruskal-Wallis and Mann-Whitney nonparametric tests, as implemented by the 3S program, were used to compare the coding-sequence lengths in the different compositional groups. The frequencies of long coding sequences in each compositional class were compared by means of a chi-square test (4F program).

Results

Random Sequences

Before analyzing the effects of compositional constraints on natural coding-sequence lengths, we will briefly introduce some formulae to compute the theoretical stop-codon density and the expected distribution of reading-frame lengths. In a random sequence without strand bias—i.e., with the same number of occurrences of each base on each strand—where $f_C = f_G = p$, and $f_A = f_T = q$, it is apparent that the stop-codon density (i.e., the sum over all three stop-codon frequencies) may be expressed as

$$t = f_T f_A^{\,2} + 2\, f_T f_A f_G = q^2 - q^3 \qquad (1)$$

The probability that a stop codon (S) is repeated after n non–stop codons (i.e., the probability of SX_nS, where X = any of 61 non–stop codons) is then

$$P_n = (1 - t)^n t \qquad (2)$$

This expression has the form of a geometric distribution with probability t. The expected average length for reading frames, defined as DNA segments of sense codons flanked by in-phase stop codons, is then $1/t$. Equations similar to (1) and (2) can be found in other works (Senapathy 1986; Stoltzfus et al. 1995). According to equation (2), the most frequent reading frame is one of length zero (when two stop codons occur next to each other). Also, the shorter the reading frame, the more frequently it appears. As n increases, the probability decreases exponentially, and thus the longer the reading frame, the rarer it becomes. The negative exponential distribution of reading frames is upper unbounded—P_n decreases smoothly and monotonically toward, but never reaching, zero. Therefore, there is no threshold or maximum length limit for reading-frame lengths (see Stoltzfus et al. 1995). In DNA sequences, the expected distribution is somewhat altered due to a lower size limit for exons (Höglund et al. 1990; see also Stoltzfus et al. 1995).

The main spatial variation along DNA sequences is the fluctuation of the AT/GC ratio, leading to a compositional heterogeneity with many important biological consequences (Bernardi et al. 1985; Bernardi 1989, 1993, 1995; Holmsquist 1989). Therefore, although other compositional biases, such as the variation of the AG/CT ratio (strand bias), may also play a role, we will focus here mainly on the broader effects of GC content variations.

Figure 1 shows the reading-frame distributions expected in three random sequences of different GC contents. The variation in the expected average reading-frame length (computed as $1/t$) with sequence composition is shown in Fig. 2, the bounds of GC content having been set to the physiological values 25% and 75%. Given the exponential function, the reading frame lengthens incrementally with rising GC levels. In fact, the expected reading-frame length doubles as the GC level goes from about 35% to around 60%, corresponding, respectively, to the lowest and the highest GC levels in the warm-blooded vertebrates analyzed.

DNA Sequences

We first look for the effects of interspecies compositional variations. The relationship between GC content

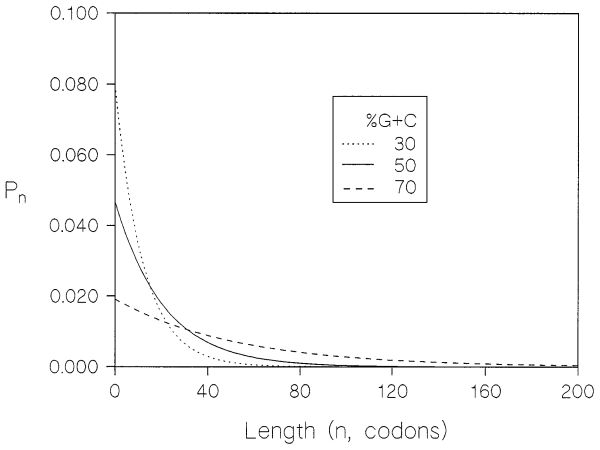

Fig. 1. Probability distributions of reading-frame lengths (computed by equations 1 and 2) in random sequences of three different GC contents (30, 50, and 70%).

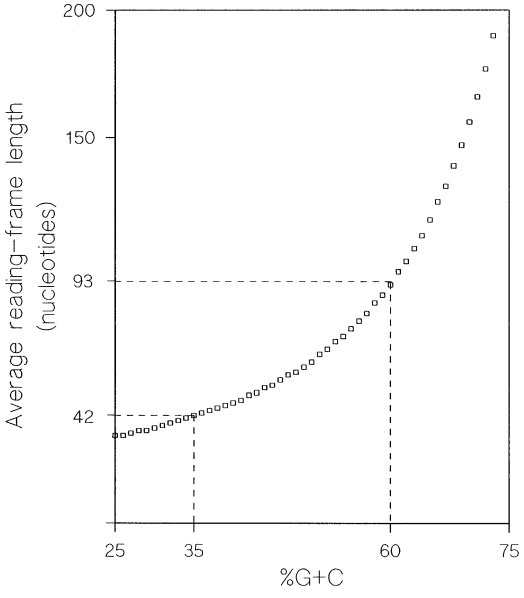

Fig. 2. Expected average reading-frame lengths (computed as 1/*t*) in random sequences of different GC contents. Each point represents the average length of the distribution of reading frames corresponding to a given G+C content.

and CDS length was sought by analyzing two groups of coding sequences: the contiguous prokaryotic genes and the exons of split genes from the vertebrates. The average coding-sequence lengths observed in each genome are given in Table 1. The expected average reading-frame length, computed as 1/*t* according to the average GC content, is also provided for comparison. Concerning prokaryotic genomes, the expected increase in the average coding-sequence length with increasing GC content appears evident as far as *M. genitalium* is ignored; thus, the strong intergenomic variation in the GC content of the three remaining prokaryotic species provokes significant differences in the average coding-sequence lengths (Kruskal-Wallis H = 63.73, $P < 0.0001$). The greater average coding-sequence length found in *M. genitalium*

Table 1. Average lengths of genes (prokaryotes) and exons (vertebrates)[a]

Genome	*N*	Average %G+C	Average coding-sequence length		P_n
			Observed	Expected	
E. coli	1634	52.1	1061	68	0.49×10^{-8}
B. subtilis	1085	43.9	972	53	0.36×10^{-9}
H. influenzae	1726	38.5	910	46	0.80×10^{-10}
M. genitalium	468	31.6	1095	51	0.16×10^{-10}
Human	1740	54.5	162	76	0.44×10^{-2}
Cow	148	53.9	162	73	0.43×10^{-2}
Mouse	934	54.8	167	76	0.41×10^{-2}
Rat	507	53.2	158	71	0.43×10^{-2}
Chicken	399	53.2	142	71	0.56×10^{-2}

[a] The observed lengths are compared to the expected average reading-frame lengths in random sequences with the average GC content found in each genome. The expected average length in *M. genitalium* was computed taking into account the presence of only two stop codons (UAA and UAG) in this genome. The probability P_n (computed by equation 2) of a reading frame of similar length and composition as that observed in each species is shown in the last column.

could be a consequence of the fact that only two stop codons (TAA and TAG) are present in this genome. With regard to vertebrate genomes, all mammals show average GC contents and coding-sequence lengths of the same magnitude; the lower value found in the chicken could be related to the compaction phenomenon which seems to have affected this genome (see below).

A feature common to all the genomes analyzed is the departure of the observed average from the expected average of coding-sequence length; such departure is stronger in the prokaryotic genomes, where the ratio of observed to expected varies from 15.6 in *E. coli* to 21.5 in *M. genitalium*, than in the vertebrate genomes, where such quotient is approximately 2.

The last column in Table 1 lists the probabilities (P_n) of reading frames with a length and composition similar to the average in each genome. The probabilities for the extremely long prokaryotic genes are six to eight orders of magnitude lower than those obtained for vertebrate exons.

Next, we investigate the effects of intragenomic compositional variations. Within each genome, sequences were classified into three compositional groups (see Data and Methods) and the average CDS length per group was determined. Most of the differences in average coding-sequence length among the three GC classes from each genome were found to be statistically significant by means of the Kruskal-Wallis nonparametric test (not shown). Then all pairwise comparisons were made with the Mann-Whitney nonparametric test—those between the two more extreme compositional classes, GC-poor vs GC-rich coding sequences, are presented in Tables 2 and 3. With only the exception of the chicken among the

Table 2. Average exon length and long-exon frequency in five eukaryotic species[a]

Genome	Compositional class	Number of exons	Average %G+C ± SE	Exon length: Average ± SE	Exon length: Comparison GC-poor/GC-rich *P*	Long exon: Frequency	Long exon: Comparison GC-poor/GC-rich *P*
	GC-poor	580	43.7 ± 0.2	134 ± 6		24.6	
Human	GC-medium	578	55.5 ± 0.1	149 ± 5	< 0.0001	30.6	< 0.0001
	GC-rich	582	64.3 ± 0.2	201 ± 13		43.0	
	GC-poor	50	42.3 ± 0.8	82 ± 9		12.0	
Cow	GC-medium	49	53.7 ± 0.5	147 ± 16	< 0.0001	30.6	< 0.0001
	GC-rich	49	65.8 ± 0.7	258 ± 62		53.1	
	GC-poor	311	46.3 ± 0.3	142 ± 9		23.5	
Mouse	GC-medium	312	55.0 ± 0.1	170 ± 8	< 0.05	39.4	< 0.002
	GC-rich	311	63.2 ± 0.2	190 ± 12		35.4	
	GC-poor	169	45.7 ± 0.3	136 ± 9		27.2	
Rat	GC-medium	169	53.8 ± 0.1	153 ± 11	< 0.0001	30.2	< 0.0003
	GC-rich	169	60.2 ± 0.3	184 ± 11		46.7	
	GC-poor	132	43.7 ± 0.4	131 ± 7		25.7	
Chicken	GC-medium	134	52.5 ± 0.2	142 ± 13	0.64	23.9	0.29
	GC-rich	133	63.3 ± 0.5	152 ± 14		31.6	

[a] Exons were classified in three compositional classes (see Data and Methods). The frequency of long exons is the frequency of exons larger than 160 bp (the average exon length in the eukaryotic sample). The comparison between the average exon lengths in GC-poor *vs* GC-rich exon classes was made by means of the Mann-Whitney nonparametric test. The frequencies of long exons were compared by means of a chi-square test.

vertebrates and *M. genitalium* among the prokaryotes, GC-rich coding sequences are significantly longer than the poorest ones.

Another way to demonstrate the relationship between GC content and CDS length is to compare the frequency of long coding sequences in the different compositional groups. Exons longer than the average for the vertebrate sample (160 bp) were classified as long and the remaining ones as short. For prokaryotes, the cutpoints were the average lengths in each genome. The frequencies of long exons (Table 2), and the frequencies of long genes (Table 3), in GC-poor and GC-rich compositional classes were then compared by means of a chi-square test. Again, with the lack of statistical significance for the chicken and *M. genitalium*, the frequencies of long coding sequences were higher in the GC-richest class.

Just the inverse relationship was found for vertebrate introns—GC-rich introns were significantly shorter than the poorest ones in all genomes analyzed (Table 4). A cutpoint of 736 bp (the mean intron length in the vertebrate sample) was used to determine the frequencies of long introns in the different compositional classes. GC-

Table 3. Average gene length and long-gene frequency in prokaryotic genomes[a]

Genome	Compositional class	Number of genes	Average %G+C ± SE	Gene length: Average ± SE	Gene length: Comparison GC-poor/GC-rich *P*	Long gene: Frequency	Long gene: Comparison GC-poor/GC-rich *P*
	GC-poor	544	47.7 ± 0.2	810 ± 23		26.8	
E. coli	GC-medium	544	52.7 ± 0.0	1097 ± 26	< 0.0001	44.9	< 0.0001
	GC-rich	546	55.8 ± 0.1	1275 ± 33		53.3	
	GC-poor	361	39.5 ± 0.2	667 ± 23		17.5	
B. subtilis	GC-medium	361	44.3 ± 0.1	1033 ± 40	< 0.0001	40.4	< 0.0001
	GC-rich	363	47.9 ± 0.1	1214 ± 69		53.7	
	GC-poor	576	34.6 ± 0.1	745 ± 23		25.3	
H. influenzae	GC-medium	574	38.7 ± 0.0	979 ± 24	< 0.0001	48.3	< 0.0001
	GC-rich	576	42.1 ± 0.1	1006 ± 27		48.3	
	GC-poor	156	27.8 ± 0.1	1017 ± 51		32.7	
M. genitalium	GC-medium	156	31.4 ± 0.1	1186 ± 61	0.61	46.2	0.63
	GC-rich	156	35.5 ± 0.2	1082 ± 70		35.3	

[a] The gene sample from each genome was stratified in three compositional classes (see Data and Methods). The frequency of long genes is the frequency of genes larger than the average gene length in each genome. The comparison between the average gene lengths in GC-poor *vs* GC-rich gene classes was made by means of the Mann-Whitney non-parametric test. The frequencies of long genes were compared by means of a chi-square test.

Table 4. Average intron length and long-intron frequency in five eukaryotic species[a]

Genome	Compositional class	Number of introns	Average %G+C ± SE	Intron length: Average ± SE	Intron length: Comparison L1 + L2/H3 *P*	Long intron: Frequency	Long intron: Comparison L1 + L2/H3 *P*
	GC-poor	478	38.7 ± 0.3	1440 ± 179		50.6	
Human	GC-medium	477	52.8 ± 0.1	946 ± 57	< 0.0001	41.5	< 0.0001
	GC-rich	478	64.9 ± 0.3	318 ± 17		9.6	
	GC-poor	39	30.6 ± 0.7	711 ± 98		38.5	
Cow	GC-medium	38	47.5 ± 1.1	919 ± 135	< 0.0001	39.5	< 0.0002
	GC-rich	39	66.2 ± 1.0	187 ± 33		2.6	
	GC-poor	253	43.1 ± 0.3	828 ± 49		42.7	
Mouse	GC-medium	252	51.4 ± 0.1	651 ± 78	< 0.0001	25.4	< 0.0001
	GC-rich	253	58.7 ± 0.3	289 ± 24		8.7	
	GC-poor	135	42.6 ± 0.3	968 ± 73		51.1	
Rat	GC-medium	136	51.1 ± 0.1	786 ± 83	< 0.0001	34.6	< 0.0001
	GC-rich	135	58.5 ± 0.3	317 ± 50		6.7	
	GC-poor	117	37.1 ± 0.4	500 ± 36		26.5	
Chicken	GC-medium	116	46.9 ± 0.2	550 ± 70	< 0.002	19.8	< 0.03
	GC-rich	117	62.8 ± 0.8	386 ± 39		14.5	

[a] Introns were classified in three compositional classes (see Data and Methods). The frequency of long introns is the frequency of introns larger than 736 bp (the average intron length in the eukaryotic sample). The comparison between the average intron lengths in GC-poor *vs* GC-rich intron classes was made by means of the Mann-Whitney nonparametric test. The frequencies of long introns were compared by means of a chi-square test.

poor and GC-rich classes harbor the longer and the shorter intron lengths, respectively (Table 4).

Discussion

Previous work has shown that functional and structural constraints are involved in determining the length of coding sequences (Blake 1983, 1985; Hawkins 1988; Traut 1988). It is known that the size distributions of the gene parts (exons, introns, leader and trailer regions, etc.) are under stabilizing selection against extreme lengths (Smith 1988). Höglund et al. (1990), in analyzing the origin of exons from random reading frames, concluded that reading frames larger than 150 bp were probably selected as exons, and that a lower size limit—perhaps imposed by RNA splicing requirements or by the limited possibilities of the smaller peptides generating structural and functional specificity—exists, below which the probability of a reading frame being selected as an exon is very low. The evolution of proteins from random amino-acid sequences (White and Jacobs 1993) has also been explored. Recently, the exon/intron organization of vertebrate genes belonging to different isochore classes has been analyzed (Duret et al. 1995). Here, we present theoretical arguments as well as empirical evidence that the longest eukaryotic exons and the longest prokaryotic genes are the GC-richest ones; this means that the differential enlargement of coding sequences may be also constrained by compositional heterogeneity pervading most genomes.

Coding Sequences

Figure 1 shows that, given the compositional asymmetry of stop codons, the expected length for random reading frames is a function of sequence GC content; that is, the higher the GC content, the higher the probability for longer reading frames. The average reading-frame length in GC-rich sequences is expected to be larger than in the GC-poor ones (Fig. 2). Inter- and intragenome comparisons of coding-sequence lengths indicate that both expectations are fulfilled in most genomes. The average gene length is related to the average GC content in prokaryotic genomes (Table 1), and both the larger vertebrate exons (Table 2) and the larger prokaryotic genes (Table 3) are the GC-richest ones within every genome. We found, however, two exceptions to this rule. The first appeared in the chicken genome, in which the larger exon lengths and the higher frequency of long exons observed in the GC-rich compositional class were not significantly different from those found in the poorest one; a possible explanation is that the strong selective component in the genome compaction of birds (Hughes and Hughes 1995) may dilute the effects of compositional constraints on the lengths of chicken coding-sequences. Second, gene length and GC content were also unrelated within the genome of *M. genitalium,* where functional or structural constraints might be the main factors determining coding-sequence lengths; note also that, given the extreme AT richness of this minimal genome, a low response of gene lengths to GC-content variations is expected (see Fig. 2).

In all the remaining vertebrate and prokaryotic spe-

cies, evidence was found supporting the rule of coding-sequence enlargement with sequence GC content. Such a conclusion fits the theoretical expectation that coding-sequence length is a function of stop-codon density, as both contiguous genes and eukaryotic exons are more or less closely flanked by AT-biased stop codons (Senapathy 1988; Senapathy et al. 1990; Seidel et al. 1992).

Stop-Codon Usage and Coding-Sequence Length

The linking between GC content and CDS length mediated by stop-codon density does not appear to result in different gene lengths according to the particular ending codon. An analysis carried out in the prokaryotic gene sample revealed no significant differences in the average gene length among the three sets of genes defined by each stop codon (results not shown). The unequal usage of stop codons apparent in most genomes is not yet fully understood, since compositional effects as well as selective factors seem to be involved to different extents in the different genomes (Sharp et al. 1992; Poole et al. 1995).

Introns

Intron lengths show just the opposite trend of that found in exons—the GC-richest introns are the shorter ones (Table 4). However, the shortening of GC-rich introns cannot be completely accounted for by the enlargement of GC-rich exons, since (1) the effect of GC content is much more pronounced on the reduction of intron length than on the lengthening of exons (compare column 5 in Tables 2 and 4), which would be related to the lower selective constraints acting on intron lengths, and (2) in the vertebrate gene sample analyzed here, the average exon length at each gene was found to be *positively* correlated to the average intron length ($r = 0.13$, $P < 0.001$).

A strong allometric shortening of introns has been observed along the genome compaction of birds, indicating that DNA loss from chicken genes has occurred disproportionally in long introns (Hughes and Hughes 1995). Table 4 confirms this result; in addition, it shows that intron reduction seems to be limited to the chicken GC-poorest genome regions—the GC-poor introns from the chicken (500 bp on average) show a strong reduction with respect to the human ones (1,440 bp on average), whereas the chicken GC-rich introns show instead a small increase (386 vs 318 bp, respectively). Such asymmetrical changes in the intron lengths of the different GC classes suggest that, in addition to selective factors (Hughes and Hughes 1995), compositional constraints could also play a role in the genome compaction of birds.

Eukaryotic Gene Lengths

Given the direct relationship between sequence GC content and CDS length shown above, it is surprising to learn that long genes from vertebrates are scarce in GC-rich isochores (Duret et al. 1995). Thus, the shortest vertebrate genes but also the largest exons (Table 2) are GC-rich. Indeed, this seems to be a general feature of eukaryotes, as long genes from yeast are also GC-poor (not shown). These results contrast with the observation that long prokaryotic genes are the GC-richest ones (Table 3).

Duret et al. (1995) also observed that total intron length is lower, and gene compactness higher, in GC-rich isochores. We observed here that both long exons (Table 2) and short introns (Table 4) are the GC-richest ones. Furthermore, when the G+C at the third codon position is used to stratify the vertebrate gene sample in three compositional classes, the average number of exons, and consequently also the number of introns, is significantly lower in GC-rich genes than in the poorest ones (5.07 ± 0.27 vs 6.13 ± 0.40, respectively; Mann-Whitney $U = 27387$, $P < 0.03$). In summary, vertebrate GC-rich regions harbor the shorter lengths for entire genes, the shorter sums for both exons and introns, and the shorter individual introns, as well as the lower number for both exons and introns. The only exception to such a generalized compaction process seems to be the GC-richest exons, whose greater lengths may be related to the high GC pressure prevailing at GC-rich genome regions. The reason why exon lengthening does not result, however, in a larger eukaryotic protein length, might be the lower exon number of the genes harbored by GC-rich genome regions. All appear, therefore, as if the GC-rich 'housekeeping subgenome' (Holmquist 1989) underwent some type of streamlining process. Whether such subgenome compaction occurred through excision-biased recombination at introns (Duret et al. 1995), in a selective way as in birds (Hughes and Hughes 1995), or through some other process, is unknown at present.

For some genes, the size reduction in the GC-richest genome regions can even affect exons, despite the hindrance of the strong GC pressure characterizing these regions. This seems to have occurred in exon 2 of globins. From the 74 α-globin vertebrate genes retrieved from the EMBL nucleotide database, we found 66 which, on the basis of their G+C content, can be assigned to the GC-poor or GC-medium compositional classes; the average length for exon 2 in these genes is 222.4 ± 0.4 bp. A significantly lower average length (204.9 ± 0.1 bp; Mann-Whitney $U = 517$, $P < 0.0001$) for this exon was found in the remaining eight GC-richest α-globin genes.

Compositional Fluctuations and the Statistical Limit for Coding-Sequence Lengths

As mentioned above, since equation (2) is upper unbounded, an absolute upper bound for reading-frame lengths does not exist. However, a length of ≈200 codons has been proposed as the upper *statistical* limit for both

primordial reading frames and exons in present-day genomes (Senapathy 1986; Naora et al. 1987). Longer reading frames were considered extremely improbable by these authors because of the intervention of stop codons. Such a figure has continued to be used as a reference value in recent publications (Senapathy 1995; Stoltzfus et al. 1995). However, equations (1) and (2) make such a limit untenable. A statistical upper limit of 200 codons would work only for sequences under 50% G+C, but, given the exponential response of reading-frame lengths to compositional variations (equation 2), this limit could undergo major deviations with only minor fluctuations in the GC content. Thus, for example, with a GC content of 50%, the probability for a reading frame of 200 codons is $P = 3.17 \times 10^{-6}$; however, when the GC content of the sequence rises to 70%, this same probability level corresponds to a reading frame more than two times longer (≈450 codons). This opens the possibility that a biased nucleotide composition in the primordial soup, or simply random fluctuations in the spatial distribution of GC content along early nucleotide sequences, could provoke pronounced variations in primitive CDS lengths. Abandoning the statistical limit of 200 codons has also been proposed on other grounds (Stoltzfus et al. 1995).

Table 1 shows that prokaryotic genes are exceedingly long compared to the expected values. The extremely low probabilities for expected reading frames of similar length and composition mean that, besides selection and compositional constraints, other factors are probably involved in the enlargement of prokaryotic coding sequences. The derivation of present-day prokaryotic genes from ancestral split sequences by losing introns (Senapathy 1986; Holland and Blake 1990) would be one such factor.

Another Role for Compositional Heterogeneity

Compositional heterogeneity in warm-blooded vertebrates has been related to the banding pattern of metaphase chromosomes, DNA replication timing, codon usage, gene frequency, CpG island density, mutation rate, recombination frequency, and insertion of both interspersed repeats and retroviral sequences (see Bernardi 1989, 1993, 1995 for reviews). We can now add that compositional heterogeneity is also associated, in both prokaryotic and eukaryotic genomes, with a differential enlargement of coding sequences; the longest coding sequences—exons in vertebrates (Table 2) and genes in prokaryotes (Table 3)—are the GC-richest ones. This constitutes a new evolutionary meaning for genome compositional variations.

Acknowledgments. We are grateful to Drs. M. Ruiz-Rejón and J.P. Martínez-Camacho for critical readings. Meaningful comments and suggestions from an anonymous referee are greatly appreciated. Help with the manuscipt from David Nesbitt is also acknowledged. This work was supported by the DGICYT (PB93-1152-CO2-01/-02) of the Spanish Government.

References

Bernardi G (1989) The isochore organization of the human genome. Annu Rev Genet 23:637–661

Bernardi G (1993) The isochore organization of the human genome and its evolutionary history—a review. Gene 135:57–66

Bernardi G (1995) The human genome: organization and evolutionary story. Annu Rev Genet 29:445–476

Bernardi G, Olofsson B, Filipski J, Zerial M, Salinas J, Cuny G, Meunier-Rotival M, Rodier F (1985) The mosaic genome of warm-blooded vertebrates. Science 228:953–958

Blake C (1983) Exons—present from the beginning? Nature 306:535–537

Blake C (1985) Exons and the evolution of proteins. Int Rev Cytol 93:149–185

Boldögkoi Z, Murvai J, Fodor I (1995) G and C accumulation at silent positions of codons produces additional ORFs. Trends Genet 11: 125–126

Cebrat S, Dudek MR (1996) Generation of overlapping reading frames. Trends Genet 12:12

D'Onofrio G, Bernardi G (1992) A universal compositional correlation among codon positions. Gene 110:81–88

Duret L, Mouchiroud D, Gautier C (1995) Statistical analysis of vertebrate sequences reveals that long genes are scarce in GC-rich isochores. J Mol Evol 40:308–317

Fleischmann RD et al. (1995) Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd. Science 269:496–512

Fraser CM et al. (1995) The minimal gene complement of *Mycoplasma genitalium.* Science 270:397–403

Guigó R, Fickett JW (1995) Distinctive sequence features in protein coding, genic non-coding, and intergenic human DNA. J Mol Biol 253:51–60

Hawkins JD (1988) A survey on intron and exon lengths. Nucleic Acids Res 16:9893–9908

Höglund M, Säll T, Röhme D (1990) On the origin of coding sequences from random open reading frames. J Mol Evol 30:104–108

Holland SK, Blake CCF (1990) Proteins, exons, and molecular evolution. In: Stone EM, Schwartz RJ (eds) Intervening sequences in evolution and development. Oxford University Press, New York, p 32

Holmquist GP (1989) Evolution of chromosome bands: molecular ecology of noncoding DNA. J Mol Evol 28:469–486

Hughes AL, Hughes MK (1995) Small genomes for better flyers. Nature 377:391

Merino E, Balbás P, Puente JL, Bolivar F (1994) Antisense overlapping open reading frames in genes from bacteria to humans. Nucleic Acids Res 22:1903–1908

Naora H, Miyahara K, Curnow RN (1987) Origin of non coding DNA sequences: molecular fossils of genome evolution. Proc Natl Acad Sci USA 84:6195–6199

Nomura M, Sor F, Yamagishi M, Lawson M (1987) Heterogeneity of GC content within a single bacterial genome and its implications for evolution. Cold Spring Harb Symp Quant Biol 52:658–663

Perrière G, Gouy M, Gojobori T (1994) NRSub: a non-redundant data base for the *Bacillus subtilis* genome. Nucleic Acids Res 22:5525–5529

Poole ES, Brown CM, Tate WP (1995) The identity of the base following the stop codon determines the efficiency of *in vitro* translational termination in *Escherichia coli*. EMBO J 14:151–158

Seidel HM, Pompliano DL, Knowles JR (1992) Exons as microgenes? Science 257:1489–1490

Senapathy P (1986) Origin of eukaryotic introns: a hypothesis, based on codon distribution statistics in genes, and its implications. Proc Natl Acad Sci USA 83:2133–2137

Senapathy P (1988) Possible evolution of splice-junction signals in eukaryotic genes from stop codons. Proc Natl Acad Sci USA 85: 1129–1133

Senapathy P (1995) Introns and the origin of protein-coding genes. Science 268:1366–1367

Senapathy P, Shapiro MB, Harris NL (1990) Splice junctions, branch point sites, and exons: sequence statistics, identification, and applications to genome project. Methods Enzymol 183:252–278

Sharp PM, Burgess CJ, Lloyd AT, Mitchell KJ (1992) Selective use of termination codons and variations in codon choice. In: Hatfield DL, Lee BJ, Pirtle RM (eds) Transfer RNA in protein synthesis. CRC Press, Boca Raton, pp 398–425

Smith MW (1988) Structure of vertebrate genes: a statistical analysis implicating selection. J Mol Evol 27:45–55

Stoehr PJ, Cameron GN (1991) The EMBL data library. Nucleic Acids Res (Suppl) 19:2227–2230

Stoltzfus A, Spencer DF, Zuker M, Logsdon JM, Doolittle WF (1995) Introns and the origin of protein-coding genes (response). Science 268:1367–1369

Sueoka N (1992) Directional mutation pressure, selection constraints, and genetic equilibria. J Mol Evol 34:95–114

Traut TW (1988) Do exons code for structural or functional units in proteins? Proc Natl Acad Sci USA 85:2944–2948

Wahl R, Rice P, Rice CM, Kröger M (1994) ECD—a totally integrated database of *Escherichia coli* K12. Nucleic Acids Res 22:3450–3455

White SH, Jacobs RE (1993) The evolution of proteins from random amino acid sequences. I. Evidence from the lengthwise distribution of amino acids in modern protein sequences. J Mol Evol 36:79–95