J Mol Evol (1986) 24:39–44

Journal of
Molecular Evolution

Silent Nucleotide Substitutions and G + C Content of Some Mitochondrial and Bacterial Genes

Thomas H. Jukes and Vikas Bhushan

Space Sciences Laboratory, University of California, Berkeley, California 94720, USA

Summary. The G + C content of DNA varies widely in different organisms, especially microorganisms. This variation is accompanied by changes in the nucleotide composition of silent positions in codons. (Silent positions are defined and explained in the text.) These changes are mostly neutral or near neutral, and appear to result from mutation pressure in the direction of increasing either A + T (AT pressure) or G + C (GC pressure) content. Variations in G + C content are also accompanied by substitutions at replacement positions in codons. These substitutions produce changes in the amino acid content of homologous proteins. The examples studied were genes for 13 mitochondrial proteins in five species, and A and B genes for bacterial tryptophan synthase in four species.

In microorganisms, varying AT and GC mutational pressures, presumably resulting from shifts in the DNA polymerase system, exert strong effects on molecular evolution by changing the G + C content of DNA. These effects may be greater than those of random drift. The effects of GC pressure on silent substitutions in the systems examined are several times as great as the effects on replacement substitutions.

GC pressure is exerted on noncoding as well as coding regions in mitochondrial DNA. This is shown by the close correlation (correlation coefficient, 0.99) of the G + C content of the noncoding D loop of mitochondria with the G + C content of silent positions in the corresponding mitochondrial genes.

Key words: G + C content of DNA — Codon usage — Mitochondrial genes — Tryptophan synthase — Neutral theory — Silent changes — D loop

There is a wide range in G + C content of DNA among various organisms, especially microorganisms. Sueoka (1961) presented evidence that this content affected the amino acid content of bacterial proteins through the genetic code, which at the time was undeciphered.

Sueoka (1962) proposed that the GC content of a bacterial genome will be determined by the effective base conversion rates u (from GC to AT, AT pressure) and v (from AT to GC, GC pressure). The equilibrium GC content is $v/(u + v)$. When u/v is 3, the GC content is 25%. He pointed out that when the GC contents of DNA in two organisms differ appreciably, enzymes of identical function may have similar active sites, "but the dispensable parts of the molecule will be quite different," and that there is little variation in mean GC content among "invertebrates, vertebrates, and plants." Freese (1962) reached similar conclusions. He stated that portions of proteins "can be partially altered without any functional change" as a result of changes in the base ratio, that "most base pairs in DNA can undergo changes that have no or only an insignificant selective effect," and that "for each DNA species, one

Offprint requests to: T.H. Jukes

kind of base pair, e.g., GC, has been altered more frequently than the other one, resulting in a shift of the base ratio." When the code became known, Jukes (1965) proposed that microorganisms high in G + C had probably "evolved towards the use of coding triplets ending in G or C" and those low in G + C had "evolved towards the use of coding triplets ending in A or U." Cox and Yanofsky (1967) showed that the base ratio changed in the direction of increased G + C when the Treffers mutator gene, which favors the substitution of A·T by C·G, was introduced into *Escherichia coli,* and that the organism could flourish despite this change. King and Jukes (1969) adduced this finding in support of the neutral theory. Freese had concluded that random drift was not responsible for "the difference in the DNA base ratios of different species," but in presenting the neutral theory, Kimura (1968), in contrast to Freese, emphasized the role of genetic drift.

The recent availability of much new information on the nucleotide sequences of genes enables a more detailed examination to be made of the effects of differences in G + C content of homologous genes among various organisms. In this letter, we present comparisons among the members of a group consisting of 13 major mitochondrial genes and of A and B genes for bacterial tryptophan synthase. We have restricted our comparisons to G + C contents of silent sites and replacement sites in genes for homologous proteins in different species. The term "homologous" is used to mean "having a common evolutionary origin as evidenced by function and primary structure of the protein." Silent sites can undergo transitions, and in some cases transversions, without the amino acid specified by the codon changing. All other sites are replacement sites. We have used total silent sites rather than only third-codon-position sites because silent sites occur in the first positions of leucine and arginine codons. The third positions of methionine and tryptophan codons are not silent in the universal code.

Nucleotide sites subject to silent changes ("silent sites") are calculated as follows for the universal code [N = A, C, G, or T(U); R = A or G; Y = C or T(U)]: A, third positions of all codons, plus A in first positions of AGR codons; C, third positions of all codons, plus C in first positions of CTR and CGR codons; G, third positions of all codons, minus G in third positions of ATG and TGG codons; T, third positions of all codons, plus T in first positions of TTR codons.

For the vertebrate and *Drosophila* mitochondrial codes, ATR and UGR are codons for methionine and tryptophan, respectively. AGR codons are chain terminators in the vertebrate mitochondrial code, while AGA is a serine codon in *Drosophila* mitochondria. Thus, the calculation of silent sites for mitochondrial codes is simpler: A, third positions; C, third positions, plus C in first positions of CTR codons; G, third positions; T, third positions, plus T in third positions of TTR codons.

Stop codons are excluded from all calculations.

The comparison is summarized in Tables 1 and 2. Most of the "adjustment" to base-ratio pressure takes place in the silent sites. This was noted for third codon positions in *Drosophila yakuba* mitochondrial genes by Clary and Wolstenholme (1985), in *Pseudomonas aeruginosa* trp G + A + B genes by Crawford (personal communication and manuscript in preparation), and in *Thermus thermophilus* by Kagawa et al. (1984).

Table 1 shows trends in amino acid composition by comparing percentages of codons with A and T in the first two positions with those of codons with G and C in the first two positions. TTR leucine codons are not included in the calculations because T in the first position is silent, and CGR arginine codons are not included in the calculations for trp genes for the same reason. The trends indicate the "evolutionary strategy" for adjusting to marked disparities in G + C content; the adjustments extend beyond use of substitutions in silent positions.

In mitochondrial genes of *D. yakuba* as compared with those of vertebrates, low G + C (23.4%) is accompanied by increasing phenylalanine, asparagine, and tyrosine, but not by increasing isoleucine, lysine, or methionine. Simultaneously, there are decreases in alanine and proline, but not in glycine or arginine. In the case of trp genes in bacteria, high G + C (68.1%) is accompanied by decreases in amino acids coded for by codons with A and T in the first two positions, especially lysine, and by simultaneous increases in codons with G and C in the first two positions, especially alanine, arginine (CGN codons), and glycine. The "responses" in mitochondria are different from those in trp genes, probably reflecting differences in protein structure and function.

The differences in Table 1 for specific, homologous proteins are very similar to the original observations by Sueoka (1961) that high A + T content of DNA in bacteria was correlated with increases in isoleucine, lysine, phenylalanine, and tyrosine in total proteins, and high G + C content with increases in alanine, arginine, and glycine.

The G + C content of silent nucleotide sites is compared with the total G + C content in two sets of homologous genes in Table 2. The G + C content of silent sites of all genes examined varies over a much wider range (5.8–92.2%) than that of replacement sites (34.5–55.0%), reflecting the fact that amino acid replacements, because of constraints, take place less frequently than substitutions at silent sites. This is evident in Tables 1 and 2 and in Fig. 1.

Table 1. Percentages of "AT" and "GC" codons for certain amino acids

	Mitochondrial genes (13) (percentage of amino acids)					trp A and B genes (percentage of amino acids)			
Amino acids	*D. yak.*	*X. laev.*	Mouse	Bovine	Human	*B. sub.*	*E. coli*	*S. typh.*	*Ps. aerug.*
Number of codons examined	(3735)	(3751)	(3819)	(3798)	(3789)	(664)	(677)	(667)	(668)
"AT"									
Phe	8.8	6.1	6.3	6.3	5.8	4.1	3.7	3.8	3.4
Ile	9.6	9.0	9.8	8.6	8.5	6.5	6.2	6.3	5.5
Met	5.7	5.2	6.5	6.6	5.7	2.2	3.1	3.0	2.7
Tyr	4.6	3.1	3.2	3.6	3.6	3.9	2.8	2.8	2.5
Asn	5.5	4.0	4.4	4.3	4.3	3.6	3.4	3.3	2.7
Lys	2.3	2.6	2.7	2.7	2.5	6.8	4.6	4.0	3.3
(a) Total	36.5	30.1	32.9	32.1	30.4	27.1	23.8	23.3	20.2
"GC"									
Pro	3.5	5.4	5.4	5.1	5.8	4.4	5.5	5.3	4.5
Ala	4.6	6.3	6.1	6.5	6.7	3.6	13.4	12.5	11.8
Arg (CG)	1.6	1.8	1.7	1.7	1.7	1.4	3.6	4.3	4.9
Gly	5.9	5.9	5.6	5.8	5.6	3.1	9.2	9.5	10.5
(b) Total	15.6	19.4	18.8	19.0	19.8	22.4	31.7	31.5	31.7
(b)/(a)	0.43	0.64	0.57	0.59	0.62	0.83	1.30	1.36	1.57

D. yak., *Drosophila yakuba*; *X. laev.*, *Xenopus laevis*; *B. sub.*, *Bacillus subtilis*; *E. coli*, *Escherichia coli*; *S. typh.*, *Salmonella typhimurium*; *Ps. aerug.*, *Pseudomonas aeruginosa*. Data for mitochondrial genes from Anderson et al. (1981, 1982), Bibb et al. (1981), Roe et al. (1985), and Clary and Wolstenholme (1985); data for trp genes from Crawford et al. (1980), Hadero and Crawford (1986), and Henner et al. (1984). "AT" and "GC" refer to codons with AA, AT, TA, and TT, or GG, GC, CG, and CC, respectively, in the first two nucleotide positions. "Arg (CG)" percentages are for CGN codons in mitochondrial genes and for CGY codons in trp genes (see text)

In comparing the nucleotide compositions of silent and nonsilent sites and the codon contents of genes, we used two homologous sets. We did not compare nonhomologous genes from different organisms. We included both animal mitochondrial genes and bacterial trp genes so as to cover a wide range of G + C contents. Our comparisons are limited to bacteria and mitochondria, both of which have relatively small amounts of noncoding DNA compared with eukaryotes.

As a working hypothesis, let us consider that changes in base ratio result from changes in the DNA polymerase system, as in the example of the Treffers mutator gene. The wobble rules of codon–anticodon pairing provide a simple mechanism for shifts in the T/C (U/C) ratio of third-codon-position nucleotides, because G in the first anticodon position can pair with either U or C.

Clary and Wolstenholme (1985) noted that *D. yakuba* mitochondrial DNA contained 92.8% A + T (7.2% G + C) in a noncoding section, the D loop. We compared this D loop with the corresponding sections of DNA in four vertebrate mitochondria. The results show close correspondence (high statistical significance) between the G + C content in these noncoding sections and the G + C content in silent positions in the 13 coding genes (Table 2 and Fig. 2). If we assume that nucleotide substitutions in D loops resulting from GC pressure (presumably imposed by the DNA polymerase system) are neutral or near neutral, then it would follow that such nucleotide substitutions in silent sites are also neutral or near neutral. The G + C content of the D loop, which is not translated, is not influenced by codon selection, but in spite of this it has almost the same G + C content as do the silent sites in the coding regions.

There are eight GNN anticodons in the vertebrate and *Drosophila* mitochondrial codes (Anderson 1982; Clary and Wolstenholme 1985). Each of these pairs with codons from a corresponding two-codon NNY set. These are for Phe, Ile, Tyr, His, Asn, Asp, Cys, and Ser (AGY codons). Four-codon sets for single amino acids have one UNN anticodon per set, so that U can pair with either U or C in third codon positions. This shift from NNC to NNT codons appears to occur readily in mitochondrial genes, as mentioned by Roe and coworkers (1985). In *D. yakuba*, the C/T ratio in third-position sites is 1:15, and in the corresponding genes in human mitochondria, the ratio is 1:0.40. For all amino acids with NNY codons, the difference between U and C usage in third codon positions in *D. yakuba* is 48.4 − 3.3 = 45.1% and in human it is 16.5 − 41.4 =

Table 2. Comparison of base compositions and silent nucleotide sites in mitochondrial (M) and trp A + B (TS) genes

	Nucleotide					Percentage G + C
Species and gene	A	C	G	T	Total	
Drosophila yakuba M						
Total	3,617	1,259	1,372	4,978	11,226	23.4
Silent sites	1,696	145	107	2,375	4,323	5.8
Replacement sites	1,921	1,114	1,265	2,603	6,903	34.5
Xenopus laevis M						
Total	3,483	2,626	1,484	3,660	11,253	36.5
Silent sites	1,623	957	186	1,373	4,139	27.6
Replacement sites	1,860	1,669	1,298	2,287	7,114	41.7
Mouse M						
Total	3,734	2,373	1,406	3,444	11,457	37.4
Silent sites	1,743	1,291	157	1,067	4,258	34.0
Replacement sites	1,991	1,582	1,249	2,377	7,199	39.3
Bovine M						
Total	3,589	3,046	1,495	3,264	11,394	39.8
Silent sites	1,625	1,483	211	919	4,012	40.0
Replacement sites	1,965	1,583	1,284	2,345	7,382	40.0
Human M						
Total	3,277	3,635	1,485	2,970	11,367	45.0
Silent sites	1,353	1,890	241	715	4,199	50.8
Replacement sites	1,924	1,745	1,245	2,255	7,168	41.7
Bacillus subtilis TS						
Total	578	395	499	520	1,992	44.9
Silent sites	184	151	133	234	702	40.5
Replacement sites	394	244	366	286	1,290	47.3
Escherichia coli TS						
Total	467	553	564	447	2,031	55.0
Silent sites	115	255	170	179	719	59.1
Replacement sites	352	298	394	268	1,312	52.7
Salmonella typhimurium TS						
Total	435	576	561	435	2,007	56.6
Silent sites	97	276	169	159	701	63.5
Replacement sites	338	390	392	276	1,306	53.0
Pseudomonas aeruginosa TS						
Total	364	730	635	275	2,004	68.1
Silent sites	33	445	205	22	705	92.2
Replacement sites	331	285	430	253	1,299	55.0

−24.9%, for a net difference between the species of 70%. The ratio is further widened by the switch from UUR to CUR as leucine codons. Simultaneously, there is little or no change in usage of A and G in silent (in this case third) codon positions (Table 2). G is used only seldom in third codon positions in any of the mitochondrial examples (range, 2.5–5.8%), so there is little possibility for change in the G content of third codon positions. The change in base ratio in mitochondrial genes evidently results from replacement of C by T progressively from human to *Drosophila*, especially in silent sites, as shown in Table 2.

Figure 1, taken from the data in Table 2, shows the contrast in slopes between plotting the G + C content at silent sites and at replacement sites against the total G + C content of genes. The G + C content in silent sites can probably range between 0% and 100%, with slopes (m) of 2.1 (for mitochondria) and 2.2 (for trp synthase A + B). The mitochondrial lines intersect at 41% total G + C content of genes, which represents the point where there is no AT or GC pressure. The corresponding point for trp synthase is 50% G + C. The observed range in replacement sites is between the narrower limits of 34.5% and 55.0%, with a slope of 0.33. The amino acid composition of proteins is necessarily limited by functional requirements, but evidently the composition can be changed, within bounds, by AT or GC pressure (Table 1). Response to this pressure is about six times as great in silent sites as in replacement sites, in our examples.

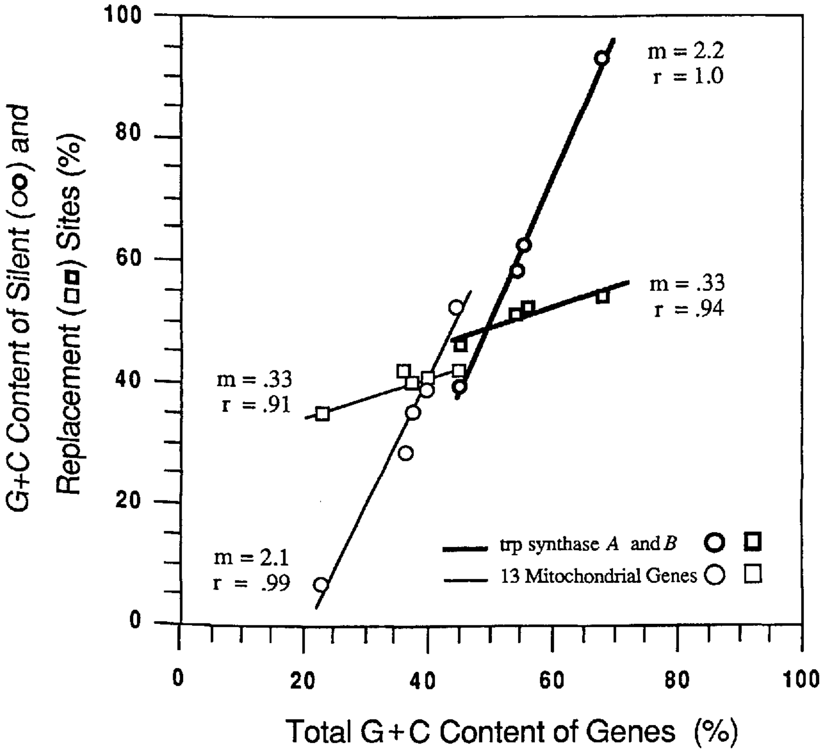

Fig. 1. G + C contents of silent (○) and replacement (□) sites plotted against total G + C contents for mitochondrial and trp synthase genes (see text).

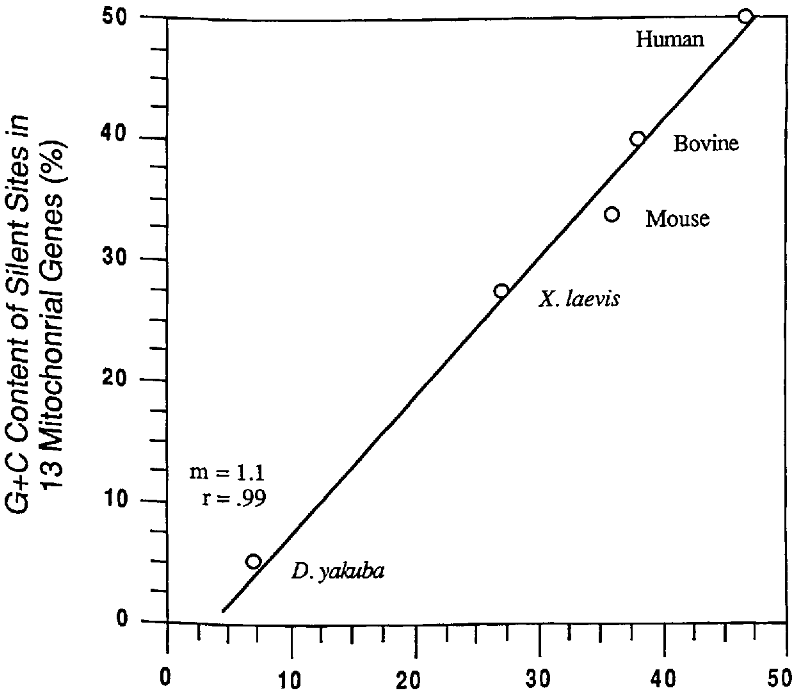

Fig. 2. G + C contents of silent sites in mitochondrial genes plotted against G + C contents of D loops (noncoding). Percentage G + C contents are as follows: *D. yakuba*, 5.6 (silent sites) and 7.2 (D loop, 1099 nucleotides); *X. laevis*, 27.6 (silent sites) and 27.3 (D loop, 2134 nucleotides); mouse, 34.0 (silent sites) and 36.6 (D loop, 879 nucleotides); bovine, 40.0 (silent sites) and 38.1 (D loop, 910 nucleotides); human, 50.8 (silent sites) and 47.1 (D loop, 1124 nucleotides)

Bernardi and Bernardi (1985) plotted G + C levels of third codon positions, rather than total silent positions, against G + C levels of corresponding genomes in many different organisms. They found a general linear relationship, but their comparisons were made between genes for different proteins. This procedure scattered their points, because in the case of nonhomologous genes from different organisms, there will be differences in amino acid content among proteins. There are also marked differences among organisms in the amounts and availabilities of tRNA species, and this affects codon choice, as shown by Ikemura (1982, 1985). Even in the same organism, there are differences in codon choice between the genes for abundant protein molecules and the genes for proteins that are made in only small amounts (Ikemura 1982, 1985). Ikemura (1985) has pointed out that codon choices of organisms with unusually high or low G + C contents "seem to be determined by the combined constraints imposed by tRNA content and by the genomic G + C content." In the case of the mitochondrial genes summarized in Tables 1 and 2, there is only one tRNA for each amino acid, plus one additional tRNA each for leucine and serine, so that the first of Ikemura's two variables is eliminated and the constraints depend solely on the G + C content. Perhaps this is also true for the bacterial trp genes summarized in Tables 1 and 2, since Ikemura (1985) has noted that tRNA populations of the Enterobacteriaceae "have been well conserved during evolution."

Kimura (1983) pointed out that the neutral theory is not antagonistic to Darwinian selection, but that it emphasizes "the much greater role of mutation pressure and random drift at the molecular level." In this letter, we have given additional emphasis and substantiation to the role of mutation pressure as originally proposed by Sueoka (1961, 1962).

Acknowledgments. Our thanks go to Irving Crawford for supplying advice, stimulation, and unpublished information, to Motoo Kimura for suggestions, to Carol Fegté for editorial assistance, and to NASA for grant NGR 05 003 460.

References

Anderson S, Bankier AT, Barrell BG, de Bruijn MHL, Coulson AR, Drouin J, Eperon IC, Nierlich DP, Roe BA, Sanger F, Schreier PH, Smith AJH, Staden R, Young IG (1981) Sequence and organization of the human mitochondrial genome. Nature 290:457–465

Anderson S, de Bruijn MHL, Coulson AR, Eperon IC, Sanger F, Young IG (1982) The complete sequence of bovine mitochondrial DNA: conserved features of the mammalian mitochondrial genome. J Mol Biol 155:683–717

Bernardi G, Bernardi G (1985) Codon usage and genome composition. J Mol Evol 22:363–365

Bibb MJ, Van Etten RA, Wright CT, Walberg MW, Clayton DA (1981) Sequence and gene organization of mouse mitochondrial DNA. Cell 26:167–180

Clary DO, Wolstenholme DR (1985) The mitochondrial DNA molecule of *Drosophila yakuba*: nucleotide sequence, gene organization, and genetic code. J Mol Evol 22:252–271

Cox EC, Yanofsky C (1967) Altered base ratios in the DNA of an *Escherichia coli* mutator strain. Proc Natl Acad Sci USA 58:1895–1902

Crawford IP, Nichols BP, Yanofsky C (1980) Nucleotide sequence of the trp B gene. J Mol Biol 142:489–502

Freese E (1962) On the evolution of the base composition of DNA. J Theor Biol 3:82–101

Hadero A, Crawford IP (1986) Nucleotide sequence of the genes for tryptophan synthase in *Pseudomonas aeruginosa*. Mol Biol Evol 3:191–204

Henner DJ, Band L, Shimotsu H (1984) Nucleotide sequence of the *Bacillus subtilis* tryptophan operon. Gene 34:169–177

Ikemura T (1982) Correlation between the abundance of yeast transfer RNAs and the occurrence of the respective codons in protein genes. J Mol Biol 158:573–597

Ikemura T (1985) Codon usage, tRNA content, and rate of synonymous substitution. In: Ohta T, Aoki K (eds) Population genetics and molecular evolution. Japan Scientific Society Press, Tokyo, and Springer-Verlag, Berlin, pp 385–406

Jukes TH (1965) The genetic code, II. Am Sci 53:477–487

Kagawa Y, Nojima H, Nukiwa N, Ishizuka M, Nakajima T, Yasuhara T, Tanaka T, Oshima T (1984) High guanine plus cytosine content in the third letter of codons of an extreme thermophile. J Biol Chem 259:2956–2960

Kimura M (1968) Evolutionary rate at the molecular level. Nature 217:624–626

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

King JL, Jukes TH (1969) Non-Darwinian evolution. Science 164:788–798

Roe BA, Ma D-P, Wilson RK, Wong JF-H (1985) The complete nucleotide sequence of the *Xenopus laevis* mitochondrial genome. J Biol Chem 260:9759–9774

Sueoka N (1961) Correlation between base composition of deoxyribonucleic acid and amino acid composition of protein. Proc Natl Acad Sci USA 47:1141–1149

Sueoka N (1962) On the genetic basis of variation and heterogeneity of DNA base composition. Proc Natl Acad Sci USA 48:582–592

Received June 6, 1986/Revised August 4, 1986