Proc. Natl. Acad. Sci. USA
Vol. 91, pp. 3799–3803, April 1994
Evolution

Pervasive CpG suppression in animal mitochondrial genomes

LON R. CARDON†, CHRIS BURGE†, DAVID A. CLAYTON‡, AND SAMUEL KARLIN†

†Department of Mathematics, Stanford University, Stanford, CA 94035; and ‡Department of Developmental Biology, Stanford University School of Medicine, Stanford, CA 94035

Contributed by Samuel Karlin, December 30, 1993

ABSTRACT **All available complete mitochondrial genomes (21 species) are evaluated for dinucleotide over- and under-representation. The CpG dinucleotide is pervasively under-represented in all animal mitochondria, but it is of variable relative abundance in fungal, protist, and plant mitochondrial genomes. Interpretations and hypotheses are considered relative to mitochondrial genome organization, methylation, structural specificities, directed mutation, and evolutionary events. In particular, our results support *Mycoplasma capricolum* or a close relative as the most likely bacterial ancestor of the mitochondria.**

Over the past two decades and continuing, mitochondrial molecular biology has been under intensive study, yielding data and insights on genome structure, organization, expression, and evolution (for recent reviews see ref. 1). The complete mitochondrial genomes of 15 animals, 3 fungi, 2 protists, and 1 plant have been sequenced. The animal mitochondrial genomes range in size from 13.8 to 17.5 kb (Table 1). Fungal and protist mitochondrial genome sizes are larger and their compositions are more variable. Plant mtDNA can be as large as bacterial genomes (2); the entire liverwort mtDNA sequence consists of 186.6 kb. Despite the size differences there is little difference in gene content. Special features of mitochondrial molecular processes and mechanisms include the following: deviations from the universal genetic code (reviewed in ref. 3); unusual tRNA structures, often missing the D arm (4, 5); almost invariant gene organization (2); trans-acting protein factors in RNA processing, especially in fungal mitochondria [e.g., "maturases" (6)]; RNA editing and trans-splicing in *Trypanosoma* mitochondria (reviewed in ref. 2); and almost strict maternal inheritance.

METHODS

Let f_X denote the frequency of mononucleotide X in the sequence at hand, f_{XY} the frequency of dinucleotide XY, and so on. A usual assessment of dinucleotide bias is through the odds-ratio measure, $\rho_{XY} = f_{XY}/f_X f_Y$. From data experience and statistical theory, for ρ_{XY} more (less) than 1.23 (0.78), the XY pair is considered to be of high (low) relative abundance compared with a random association of mononucleotides. There are statistical tests indicating the degree of statistical significance for high and low relative abundance (e.g., see refs. 7 and 8). The measure ρ_{XY} is suitable for a single sequence. To accommodate the complementary antiparallel structure of double-stranded DNA, we form the symmetrized frequency of mononucleotides as $f^*_A = f^*_T = (f_A + f_T)/2$ and $f^*_C = f^*_G = (f_C + f_G)/2$, and $f^*_{GT} = (f_{GT} + f_{AC})/2$ is the symmetrized double-stranded frequency of GT·AC, etc. The symmetrized dinucleotide odds ratio measure is taken to be $\rho^*_{GT} = f^*_{GT}/f^*_G f^*_T = 2(f_{GT} + f_{AC})/(f_G + f_C)(f_T + f_A)$ and similarly for all dinucleotides (see ref. 9 for rationale and justifications). A third-order measure of similar type is $\gamma^*_{XYZ} = f^*_{XYZ} f^*_X f^*_Y f^*_Z / f^*_{XY} f^*_{YZ} f^*_{XNZ}$ where N is any nucleotide. Higher-order measures based on tetranucleotide (or longer oligonucleotides) are also available (9).

The mononucleotide frequencies in animal mtDNA sequences tend to be asymmetric in A vs. T and C vs. G in their heavy and light strands (Table 1). Consequently, we evaluate both single-stranded (ρ) and symmetrized double-stranded (ρ^*) representation measures in the mtDNA sequences.

RESULTS

CpG Deficiencies. *Without exception all animal (vertebrate and invertebrate) mitochondrial sequences show significant CpG suppression* (relative under-abundance; Table 1), *almost to the same extent as occurs in vertebrate nuclear genomic sequences*. The traditional explanation ascribed to the methylation-deamination-mutation scenario (CpG → TpG·CpA mutations) with concomitant excess of TpG·CpA dinucleotides may not apply to mtDNA genomes, since the associated methylase is absent from most invertebrate hosts (e.g., *Drosophila, C. elegans, Strongylocentrotus purpuratus*) or the methylase does not or cannot access the mitochondrial organelle (in vertebrates). Most experiments have been unable to detect methylase activity in vertebrate mtDNA (10, 11), although some ambiguous results in this respect have been reported (e.g., ref. 12). Furthermore, contrary to expectations under CpG methylation, the TpG·CpA product of methylase mutations is not over-represented in mitochondrial sequences.

The single persistent significant relative over-abundance among dinucleotides occurs for CpC·GpG in animal mitochondrial sequences (Table 1). In vertebrates, this overabundance is mainly attributable to high representations of GpG doublets (all single-strand ρ_{GG} values > 1.35) with normal abundances of CpC doublets ($1.05 \leq \rho_{CC} \leq 1.23$). The asymmetric composition of these mitochondrial genomes indicates that although there are relatively few guanine nucleotides, they often appear in tandem arrays. The invertebrate, fungal, protist, and plant mitochondria are more symmetric in base composition and show high ρ representation values for both GpG and CpC dinucleotides.

Fungal, Protist, and Plant CpG Relative Abundances. *In the fungal group only* Schizosaccharomyces pombe *mtDNA shows significant CpG under-abundance*. Neither of the larger fungal mitochondrial genomes, *S. cerevisiae* or *Podospora anserina*, yields low CpG representations. For *S. cerevisiae*, $\rho^*_{CG} = 1.49$ is inordinately high. This can be accounted for by virtue of more than 100 separate C+G-rich clusters (each about 50 bp long); the intergenic spacer regions are A+T-rich. This anomalous nucleotide distribution in *S. cerevisiae* mtDNA yields several dinucleotide extremes (not shown).

We did not have available a complete mt genome for *Neurospora crassa*. On the basis of partial sequences totalling 35.5 kb, the lowest dinucleotide relative abundance occurred for CpG ($\rho^*_{CG} = 0.75$) and the highest for CpC·GpG

The publication costs of this article were defrayed in part by page charge payment. This article must therefore be hereby marked "*advertisement*" in accordance with 18 U.S.C. §1734 solely to indicate this fact.

Table 1. Relative abundances of selected dinucleotides in complete mitochondrial genomes

Organism	Length, bp	Base composition, mol %					TpA		TpG·CpA[a]			CpC·GpG[a]			GpC		CpG		Host (ρ^*) CpG
		A	C	G	T	C+G	ρ^*	ρ	ρ^*	ρ	ρ	ρ^*	ρ	ρ	ρ^*	ρ	ρ^*	ρ	
Vertebrates																			
Human	16,569	31	31	13	25	44	1.07	1.09	1.00	0.96	0.96	1.35	1.09	1.50	0.87	1.04	**0.53**	**0.64**	0.42
Mouse	16,295	35	24	12	29	36	1.03	1.04	1.00	0.92	1.00	1.36	1.14	1.59	0.94	1.06	**0.52**	**0.58**	0.36
Rat	16,298	34	26	12	27	38	1.01	1.02	1.00	0.92	0.98	1.39	1.16	1.59	0.88	1.01	**0.53**	**0.60**	
Cow	16,338	33	26	13	27	39	1.07	1.08	0.98	0.93	0.96	1.31	1.13	1.43	0.96	1.07	**0.56**	**0.62**	0.48
Harbor seal	16,826	33	27	14	25	41	1.09	1.11	1.03	0.97	1.00	1.24	1.05	1.40	0.87	0.97	**0.65**	**0.72**	
Fin whale	16,398	33	27	13	27	40	1.07	1.08	1.00	0.94	0.98	1.31	1.10	1.45	0.92	1.04	**0.54**	**0.62**	
Carp	16,364	32	27	16	25	43	1.05	1.06	0.99	0.98	0.95	1.30	1.17	1.36	0.95	1.03	**0.62**	**0.67**	
Bonyfish[b]	16,558	29	29	17	25	46	1.05	1.05	0.96	1.01	0.91	1.36	1.23	1.41	0.94	1.01	**0.60**	**0.65**	
Chicken	16,775	30	32	14	24	46	0.99	1.01	1.03	1.06	0.95	1.37	1.11	1.52	0.82	0.99	**0.46**	**0.55**	0.51
Xenopus laevis	17,553	33	23	13	30	37	0.99	0.99	0.99	0.96	0.98	1.28	1.13	1.39	1.02	1.10	**0.63**	**0.68**	0.50
Invertebrates																			
Drosophila yakuba	16,019	39	12	9	39	21	0.95	0.95	0.92	0.88	0.94	1.67	1.53	1.85	1.34	1.37	**0.68**	**0.69**	0.97[c]
Caenorhabditis elegans	13,794	31	9	15	45	24	0.97	1.00	0.92	0.77	1.13	1.52	1.54	1.38	1.07	1.15	**0.56**	**0.60**	0.94
Ascaris suum	14,284	22	8	20	50	28	0.83	0.97	1.11	0.97	0.82	1.61	1.78	1.28	0.72	0.90	**0.36**	**0.45**	
Paracentrotus lividus	15,696	31	23	17	30	40	0.93	0.93	0.89	0.93	0.86	1.31	1.27	1.31	1.02	1.04	**0.58**	**0.59**	
Strongylocentrotus purpuratus	15,650	29	23	18	30	41	0.92	0.93	0.89	0.91	0.89	1.33	1.27	1.38	1.04	1.05	**0.56**	**0.57**	
Fungi																			
Saccharomyces cerevisiae	78,521	42	8	9	40	17	1.22	1.22	0.65	0.66	0.64	3.12	3.35	2.91	1.29	1.29	**1.48**	**1.49**	0.80
Schizosaccharomyces pombe	19,431	34	14	16	36	30	0.91	0.91	0.93	0.92	0.93	1.32	1.25	1.37	0.94	0.94	**0.54**	**0.54**	0.90
Podospora anserina	100,314	36	13	17	34	30	1.06	1.06	0.84	0.85	0.85	1.25	1.22	1.24	1.29	1.30	**0.84**	**0.85**	
Protists																			
Trypanosoma brucei[d]	23,016	42	9	14	35	23	0.82	0.83	1.04	1.01	1.14	1.87	2.51	1.49	1.10	1.15	**0.58**	**0.61**	0.93
Paramecium aurelia	40,469	25	22	19	33	41	0.81	0.82	0.78	0.78	0.79	1.17	1.10	1.26	1.20	1.20	**0.84**	**0.84**	
Plants																			
Liverwort	186,608	29	21	21	29	42	0.85	0.85	0.91	0.91	0.92	1.22	1.20	1.23	1.09	1.09	**0.93**	**0.93**	

[a]Single-strand (ρ) and symmetrized (ρ^*) relative abundance values are presented for each dinucleotide. The single-strand values are listed first for the left dinucleotide and second for the right dinucleotide. Single-strand values are calculated by $\rho_{XY} = f_{XY}/f_X f_Y$, where X, Y = A, C, G, T. Symmetrized relative abundances are used to account for the double-stranded antiparallel structure of DNA: $\rho^*_{XY} = f^*_X f^*_Y / f^*_{XY}$, where $f^*_X = (f_X + f_{\bar{X}})/2$ and $f^*_{XY} = (f_{XY} + f_{(\bar{X}\bar{Y})})/2$.

[b]Organism is *Crossostoma lacustre*.

[c]Host data are from *Drosophila melanogaster*.

[d]Complete coding sequence from *Trypanosoma brucei* kinetoplast maxicircle genes, GenBank accession no. M94286.

($\rho^*_{CC} = 1.52$), consistent with the pattern observed in metazoan and *Schizosaccharomyces pombe* mitochondria.

The *Trypanosoma brucei* and *Paramecium aurelia* mitochondrial genomes also show low values for ρ^*_{CG} (significantly low only for *Trypanosoma*).

The large liverwort mtDNA genome carries normal CpG relative abundances. The same applies to the five available chloroplast genomes examined (Table 2).

CpG Deficiency in Coding Regions. CpG relative abundances at codon sites (1,2), (2,3), and (3,1) and separately for rRNA genes in animal mitochondrial genomes are reported in Table 3. Independent of the codon site pairings we observed pervasive significant CpG relative under-abundance, the lowest values almost always being at codon sites (1,2). This also corresponds to the extremely low arginine usage among coding amino acids (arginine and cysteine usages are among the three lowest amino acids in all animal mitochondria). Interestingly, *S. cerevisiae* is also substantially CpG underrepresented over coding domains.

With the exception of *Drosophila*, the animal mitochondrial rRNA genes also yield significantly low ρ^*_{CG} values.

Homogeneity of CpG Occurrences in mtDNA Genomes. We examined ρ^*_{CG} values over a sliding window of 400 bp (with 200-bp displacement). The corresponding set of ρ^*_{CG} values for the mtDNA sequences from human (median value 0.52), cow (0.58), chicken (0.52), carp (0.66), mouse (0.56), rat (0.51), seal (0.64), whale (0.55), frog (0.61), echinoderms *S. purpuratus* (0.61) and *P. lividus* (0.54), nematodes *A. suum* (0.31) and *C. elegans* (0.51), and the yeast *Schizosaccharomyces pombe* (0.49) rarely exceeded 0.90 with few exceptions (at most two or three windows), reaching maximal values of ρ^*_{CG} in the range 0.95–1.05. Thus, these mitochondrial genomes are quite uniform in low CpG relative abundance. The *D. yakuba* mitochondrial genome partitions into two halves, one exhibiting five 400-bp segments without a CpG dinucleotide, the other showing about eight 400-bp segments with $\rho^*_{CG} \geq 1.40$. Both nematodes feature a single C+G-rich cluster ($\rho^*_{CG} \geq 1.50$). In these assessments, *Trypanosoma* is

Table 2. Relative abundances of CpG in complete chloroplast genomes

Organism	Length, bp	Base composition, mol %					TpA		TpG·CpA[a]			CpC·GpG[a]			GpC		CpG	
		A	C	G	T	C+G	ρ^*	ρ	ρ^*	ρ	ρ	ρ^*	ρ	ρ	ρ^*	ρ	ρ^*	ρ
Euglena gracilis	41,017	33	11	13	43	24	0.85	0.86	0.94	0.90	0.96	1.37	1.28	1.40	1.37	1.39	**1.10**	**1.12**
Epifagus virginia	70,028	31	18	18	33	36	0.94	0.94	0.87	0.87	0.88	1.43	1.43	1.43	0.92	0.92	**0.91**	**0.91**
Rice	134,525	31	19	20	30	39	0.82	0.82	0.89	0.89	0.89	1.29	1.28	1.29	0.89	0.89	**0.86**	**0.86**
Tobacco	155,844	31	19	19	31	38	0.78	0.79	0.93	0.94	0.93	1.28	1.29	1.26	0.83	0.83	**0.87**	**0.87**
Liverwort	121,024	35	14	15	36	29	0.83	0.83	0.94	0.93	0.94	1.38	1.37	1.39	1.24	1.24	**0.87**	**0.87**

[a]As in Table 1.

Table 3. Relative abundances of CpG in mitochondrial genes and rRNA sequences

Organism	Codon position 1,2	2,3	3,1	Total coding	rRNA
Vertebrates					
Human	0.48	0.69	0.53	0.69	0.63
Mouse	0.47	0.55	0.56	0.58	0.76
Rat	0.53	0.56	0.59	0.63	0.76
Cow	0.42	0.69	0.55	0.66	0.69
Harbor seal	0.62	0.69	0.58	0.68	0.68
Fin whale	0.34	0.61	0.52	0.63	0.70
Carp	0.54	0.73	0.59	0.70	0.68
Bonyfish	0.50	0.68	0.58	0.66	0.72
Chicken	0.42	0.48	0.49	0.57	0.62
X. laevis	0.58	0.63	0.60	0.63	0.83
Invertebrates					
D. yakuba	0.44	0.84	0.84	0.59	1.80
C. elegans	0.54	0.82	0.59	0.56	0.63
A. suum	0.27	0.37	0.68	0.39	0.68
P. lividus	0.46	0.63	0.64	0.58	0.75
S. purpuratus	0.44	0.61	0.57	0.56	0.70
Fungi					
S. cerevisiae	0.83	0.82	0.64	0.67	1.49
S. pombe	0.38	0.37	0.50	0.41	0.83
P. anserina	0.73	1.03	0.75	0.65	0.98
Protists					
Trypanosoma	0.82	1.16	1.02	0.97	1.05
Paramecium	0.74	0.84	0.73	0.80	1.03
Plants					
Liverwort	0.87	0.89	0.96	0.90	0.95

also very heterogeneous, displaying 28 segments without a CpG dinucleotide and several C+G clusters.

Normal Relative Abundances of TpA Dinucleotides. The nuclear genomes of most vertebrates, plants, bacteria, phage, and viruses are significantly under-represented in TpA (9, 13, 14). In contrast, relative abundances of TpA dinucleotides in all mitochondria and chloroplast genomes are normal (close to 1).

DISCUSSION

For fungal, protist, and plant mitochondrial genomes, the status of CpG representations is variable and unpredictable. In contrast, the pervasively low relative abundance of CpG in animal mitochondrial genomes lacks a coherent explanation.

Does CpG Deficiency Relate to DNA Methylation? A connection to CpG methylation seems unconvincing, since either the relevant methylase(s) in the host does not exist (e.g., as for invertebrates) or methylase activities in the mitochondria have not been detected (in vertebrates).

Does CpG Suppression Reflect Under-representation of Specific Longer Oligonucleotides? Relative abundances of the trinucleotides GCG and CGC and of the tetranucleotides CCGG, CGCG, and GCGC are mostly normal, suggesting that CpG under-representation is not a consequence of the reduced frequencies of certain higher-order oligonucleotides.

Do CpG Deficiencies Reflect Strong Avoidance of Arginine Amino Acid Usage? This does not seem likely, since CpG suppression applies to all contiguous codon positions and to rRNA genes.

Directed Mutation and/or Selection Against CpG Dinucleotides. Is it possible that in the small mitochondrial genomes the CpG dinucleotide is more vulnerable to DNA substitutions mainly directed to CpC or GpG? The mutation rate in vertebrate mitochondrial genomes is assessed at 10-fold higher than the mutation rate in nuclear DNA but in fungal and plant mitochondrial genomes the mutation rate can be 10-fold less than for nuclear DNA. The reason for and mechanism of such a mutation bias or other selection effects, if present, are unknown.

Structural Role for CpG. The low relative numbers of CpG dinucleotides may reflect on a special structural or regulatory role. For example, CpG dinucleotides exhibit the greatest thermodynamic stability (stacking energies) of all dinucleotides (15, 16), possibly indicating a DNA structural/conformational specificity for CpG. Or is it possible that CpGs are anchor or attachment points of the chromosome to the mitochondrial inner membrane, or possibly preferred binding sites for "nucleoid" proteins, thereby permitting only limited usage?

Are CpG Dinucleotides Involved in Specific Functions in the Replication Process? The mitochondrial replication machinery is distinctive in that the heavy strand is largely synthesized first (the parental heavy strand being displaced) and the light strand is subsequently replicated (17). The A+T-rich genomic composition likely facilitates these processes by allowing the strands to disengage more easily. These processes may be more troublesome in the larger mitochondrial genomes, especially in *S. cerevisiae* with numerous C+G clusters.

Correlation of CpG Abundances and Genome Size. Inspection of Table 1 reveals that, without exception, CpG dinucleotides are significantly under-representated in mitochondrial genomes of size ≤25 kb. These genomes are almost all coding-streamlined (e.g., contain no introns) in gene organization. The other genomes, retaining the same basic gene content, are replete with group 1 and group 2 introns and with transposable insertion elements. Paralleling the mitochondrial CpG suppression, it is intriguing that in virtually *all small* eukaryotic viruses of genome size ≤30 kb (more than 75 have been completely sequenced), the CpG dinucleotide is with only a few exceptions (only some togaviruses) under-represented; $\rho^*_{CG} \leq 0.7$ and mostly $\rho^*_{CG} \approx 0.3$ to 0.6 (for details see ref. 18). In contrast, the CpG relative abundance values in all phage examined, large and small, are normal (data not shown). The range of ρ^*_{CG} values in the larger eukaryotic viruses (e.g., adenoviruses, herpesviruses, vaccinia) are variable but mostly in the normal range (18). Is there a correlation of low ρ^*_{CG} values with genome size coupled to a streamlined genomic organization?

Nuclear vs. Mitochondrial Genomic Comparisons. An inverse correlation between nuclear and mitochondrial genome architecture distinguishes animal versus fungal and protist taxa. Animal nuclear genes are rife with introns but have none in their mitochondrial genomes, whereas protist and fungal species exhibit the opposite distribution, entailing few introns in nuclear genes but many introns in their mitochondria. The higher metabolic rates in animals, concomitant to many life activities associated with cellular differentiation and coordination, may have selected for an efficient streamlined mitochondrion to supply energy needs without complications in processing mitochondrial gene products, etc. On the other hand, the reduced movements and life states with lesser energy demands in the single-celled fungal and protozoan species could permit greater flexibility in their mitochondrial composition.

Is the Low CpG Relative Abundance a Historical Remnant? The endosymbiont hypothesis generally proposes that mitochondrial genomes are derived by lateral transfer from a Gram-negative α-purple bacterium, with *Paracoccus denitrificans* a likely ancestor (2, 19). However, *P. denitrificans* shows normal CpG abundance ($\rho^*_{CG} = 1.13$), as do other α-purple bacteria, including *Rhizobium meliloti, Rhodobacter capsulatum*, and *Agrobacter tumefaciens* [ρ^*_{CG} values in these organisms are all close to 1.00 (9, 20)]. Although most common eubacteria are not CpG suppressed, there are ex-

ceptions, including the archaebacteria *Methanobacterium thermoautotrophicum* (ρ^*_{CG} = 0.576) and *Sulfolobus* (ρ^*_{CG} = 0.72), the primitive eubacterium *Thermus thermophilus* (ρ^*_{CG} = 0.749), the smallest bacterium, about 0.8 Mb genome size, *Mycoplasma capricolum* (has no cell wall and its classification is uncertain; ρ^*_{CG} = 0.730), and *Borrelia burgdorferi* (pathogen of Lyme disease; ρ^*_{CG} = 0.676). The *M. capricolum* genome is C+G-poor, about 28–33%, close to the C+G level of most animal mitochondrial genomes.

M. capricolum is particularly interesting with reference to mtDNA because it is the rare bacterium where the codon TGA is translated to tryptophan, as is universally true in mitochondrial genomes. We will argue that *M. capricolum* or a closely related bacterium is the more natural progenitor of mtDNA rather than an α-purple Gram-negative bacterium. This hypothesis was set forth by Andachi *et al.* (21) on the basis of similarities between fungal mitochondrial and *M. capricolum* tRNA sequences and structures. *M. capricolum* and mitochondria are also similar in that both are parasitic or symbiotic in eukaryotic species, have a pronounced economy in numbers of tRNA genes, and contain only one or two rRNA genes (22).

An assessment of genomic similarities between animal mitochondria and 21 diverse bacterial species (see legend to Table 4) based on *dinucleotide relative abundance distances* (20, 23) provides further strong evidence for genomic similarity between *M. capricolum* and mtDNA. These distance measures are calculated as $\delta(f, g) = \frac{1}{16}\Sigma|\rho^*_{ij}(f) - \rho^*_{ij}(g)|$ for comparison of sequences f and g, with the sum extending over all 16 dinucleotides (Table 4). They have been shown to provide meaningful and revealing comparisons in determinations of evolutionary relationships among genomic sequences (23). Applications of the dinucleotide relative abundance distances to the mitochondrial genomes versus the bacterial sequences show *M. capricolum* as among the two closest bacteria to all animal mitochondrial genomes examined [Table 4; the other close bacterium is the archaebacterium *Methanobacterium thermoautotrophicum* (I. Ladunga and S.K., unpublished work)]. As shown in Table 4, the relative sizes of $\delta(f, g)$ distance values between the animal mitochondria and *M. capricolum* correspond to moderate to weakly distant relatedness (see Table 4 legend). The mitochondrion–*M. capricolum* distances also are within the range of distances between mitochondria. By contrast, the α-purple bacteria are by a factor of at least 2 more distant from animal mitochondrial sequences than *M. capricolum* and yield $\delta(f, g)$ distance values substantially larger than those observed among the different mitochondrial genomes. The only exceptions to the striking closeness of the mitochondrion–*M. capricolum* comparisons are the mitochondria of *S. cerevisiae*, which is extreme in relation to the other mitochondria, as noted previously, and *Paramecium aurelia*, for which *M. capricolum* is the third-closest bacterium.

Perspectives. The mitochondria are special in many respects: maternal inheritance, no established DNA excision repair mechanisms, transcription-primed DNA replication, mobile elements in fungal mitochondrial genomes, etc. We might expect that the exceptional deficiency of CpG compared with GpC and all other dinucleotides is important in some capacity for these special properties. At present, there

Table 4. Distances within and between mitochondrial and bacterial genomes

	Mitochondria		Bacteria[a]			
Mitochondrion	Closest	Range $\delta(f, g)$	Closest		Second closest	
Human	Whale	0.015–0.217	*M. thermo.*	(0.104)	*M. capricolum*	(0.110)
Mouse	Rat	0.018–0.212	*M. thermo.*	(0.094)	*M. capricolum*	(0.098)
Rat	Mouse	0.018–0.214	*M. thermo.*	(0.087)	*M. capricolum*	(0.098)
Cow	Whale	0.013–0.218	*M. capricolum*	(0.097)	*M. thermo.*	(0.100)
Seal	Whale	0.050–0.228	*M. thermo.*	(0.095)	*M. capricolum*	(0.108)
Whale	Cow	0.013–0.211	*M. thermo.*	(0.103)	*M. capricolum*	(0.105)
Carp	*X. laevis*	0.025–0.202	*M. capricolum*	(0.085)	*M. thermo.*	(0.099)
Bonyfish	Carp	0.027–0.196	*M. capricolum*	(0.086)	*M. thermo.*	(0.116)
Chicken	Rat	0.029–0.187	*M. thermo.*	(0.090)	*M. capricolum*	(0.110)
X. laevis	Carp	0.025–0.215	*M. capricolum*	(0.061)	*M. thermo.*	(0.090)
D. yakuba	*C. elegans*	0.095–0.208	*M. capricolum*	(0.124)	*B. burgdorferi*	(0.148)
C. elegans	Bonyfish	0.066–0.209	*M. capricolum*	(0.110)	*Anabaena*	(0.142)
Ascaris	*Trypanosoma*	0.144–0.300	*M. capricolum*	(0.202)	*Anabaena*	(0.213)
P. lividus	*S. purpuratus*	0.014–0.223	*M. capricolum*	(0.080)	*T. thermophilus*	(0.094)
S. purpuratus	*P. lividus*	0.014–0.217	*M. capricolum*	(0.081)	*T. thermophilus*	(0.096)
S. cerevisiae	*D. yakuba*	0.376–0.526	*B. stearo.*	(0.475)	*E. coli*	(0.484)
Schizo. pombe	*P. lividus*	0.047–0.219	*M. capricolum*	(0.092)	*M. thermo.*	(0.101)
P. anserina	*X. laevis*	0.070–0.271	*M. capricolum*	(0.089)	*B. burgdorferi*	(0.113)
Trypanosoma	*C. elegans*	0.132–0.259	*M. capricolum*	(0.147)	*Anabaena*	(0.175)
Paramecium[b]	Liverwort	0.110–0.300	*T. thermophilus*	(0.107)	*B. burgdorferi*	(0.119)
Liverwort	*P. lividus*	0.087–0.224	*M. capricolum*	(0.055)	*Anabaena*	(0.077)

For sequences f and g, the distances are calculated as $\delta(f, g) = \Sigma|\rho^*_{ij}(f) - \rho^*_{ij}(g)|w_{ij}$, where $w_{ij} = \frac{1}{16}$, the sum extended over all 16 dinucleotides. $\delta(f, g)$ values are shown in parentheses. The rationale of these measures is elaborated in ref. 23. To help in the interpretations of the relative abundance distances, we give examples from major eukaryote species based on DNA collections ranging from 200 kb up to 2 Mb: closely related, δ(human, cow) = 0.041; moderately related, δ(*Drosophila, Bombyx mori*) = 0.058; weakly related, δ(human, trout) = 0.090; distantly related, δ(human, *Drosophila*) = 0.145; distant, δ(pig, *B. mori*) = 0.186; and very distant, δ(human, *E. coli*) = 0.210 (ref. 23; I. Ladunga and S.K., unpublished work).

[a]Bacterial genomes compared include all available sequences from *Agrobacterium tumefaciens, Rhizobium meliloti, Rhodobacter capsulatum, Neisseria gonorrhoeae, Pseudomonas aeruginosa, Escherichia coli, Haemophilus influenzae, Azotobacter vinelandii, Myxococcus xanthus, Bacillus subtilis, Bacillus stearothermophilus, Mycobacterium tuberculosis, Streptomyces griseus, Streptomyces lividans, Staphylococcus aureus, Mycoplasma capricolum, Borrelia burgdorferi, Anabaena* sp., *Thermus thermophilus, Halobacterium halobium,* and *Methanobacterium thermoautotrophicum.* Natural habitats of these bacteria include soil, water, and mammalian tissues. Several are pathogenic in humans and/or plants; several fix nitrogen. Further detailed features of these bacterial sequence collections are given in Karlin and Cardon (20).

[b]*M. capricolum* is the third-closest bacterium (0.122).

are no obvious clues in the known regulatory sequences for replication and gene expression that would require CpG suppression (24). Experiments aimed to increase the density or change the positions of CpG dinucleotides in animal mitochondrial genomes and evaluation of resulting viabilities or other consequences on mitochondrial mechanisms and processes could help clarify the nature of CpG suppression in mtDNA.

We gratefully acknowledge the helpful comments of Drs. B. Edwin Blaisdell, V. Brendel, and A. M. Campbell.

1. Wolstenholme, D. R. & Jeon, K. W., eds. (1992) *International Review of Cytology: Vol. 141, Mitochondrial Genomes* (Academic, San Diego).
2. Gray, M. W. (1992) *Int. Rev. Cytol.* **141,** 233–357.
3. Osawa, S., Jukes, T. H., Watanabe, K. & Muto, A. (1992) *Microbiol. Rev.* **56,** 229–264.
4. Lazowska, J., Jacq, C. & Slonimski, P. P. (1980) *Cell* **22,** 333–348.
5. Wolstenholme, D. R., Macfarlane, J. L., Okimoto, R., Clary, D. O. & Wahleithner, J. A. (1987) *Proc. Natl. Acad. Sci. USA* **84,** 1324–1328.
6. Lambowitz, A. M. & Perlman, P. S. (1990) *Trends Biochem. Sci.* **15,** 440–444.
7. Bishop, Y. M. M., Fienberg, S. E. & Holland, P. W. (1975) *Discrete Multivariate Analysis: Theory and Practice* (MIT Press, Cambridge, MA).
8. Hollander, M. & Wolfe, D. A. (1973) *Nonparametric Statistical Methods* (Wiley, New York).
9. Burge, C., Campbell, A. M. & Karlin, S. (1992) *Proc. Natl. Acad. Sci. USA* **89,** 1358–1362.
10. Dawid, I. G. (1974) *Science* **184,** 80–81.
11. Groot, G. S. P. & Kroon, A. M. (1979) *Biochim. Biophys. Acta* **564,** 355–357.
12. Pollack, Y., Kasir, J., Shemer, R., Metzger, S. & Szyg, M. (1984) *Nucleic Acids Res.* **12,** 4811–4824.
13. Nussinov, R. (1981) *J. Biol. Chem.* **256,** 8458–8462.
14. Beutler, E., Gelbart, T., Han, J., Kozieol, J. A. & Beutler, B. (1989) *Proc. Natl. Acad. Sci. USA* **86,** 192–196.
15. Breslauer, K. J., Frank, R., Blöcker, H. & Marky, L. A. (1986) *Proc. Natl. Acad. Sci. USA* **83,** 3746–3750.
16. Delcourt, S. G. & Blake, R. D. (1991) *J. Biol. Chem.* **266,** 15160–15169.
17. Clayton, D. A. (1982) *Cell* **28,** 693–705.
18. Karlin, S., Doerfler, W. & Cardon, L. R. (1994) *J. Virol.*, in press.
19. John, P. & Whatley, F. R. (1975) *Nature (London)* **254,** 495–498.
20. Karlin, S. & Cardon, L. R. (1994) *Annu. Rev. Microbiol.*, in press.
21. Andachi, Y., Yamao, F., Muto, A. & Osawa, S. (1989) *J. Mol. Biol.* **209,** 37–54.
22. Sawada, M., Muto, A., Iwami, M., Yamao, F. & Osawa, S. (1984) *Mol. Gen. Genet.* **196,** 311–316.
23. Karlin, S., Mocarski, E. S. & Schachtel, G. A. (1994) *J. Virol.* **68,** 1886–1902.
24. Shadel, G. S. & Clayton, D. A. (1993) *J. Biol. Chem.* **268,** 16083–16086.