Proc. Natl. Acad. Sci. USA
Vol. 91, pp. 12842–12846, December 1994
Evolution

Which bacterium is the ancestor of the animal mitochondrial genome?

SAMUEL KARLIN[†] AND ALLAN M. CAMPBELL[‡]

Departments of [†]Mathematics and [‡]Biological Sciences, Stanford University, Stanford, CA 94305

Contributed by Allan M. Campbell, August 18, 1994

ABSTRACT We present considerable data supporting the hypothesis that a *Sulfolobus*- or *Mycoplasma*-like endosymbiont, rather than an α-proteobacterium, is the ancestor of *animal* mitochondrial genomes. This hypothesis is based on pronounced similarities in oligonucleotide relative abundance extremes common to animal mtDNA, *Sulfolobus*, and *Mycoplasma capricolum* and pronounced discrepancies of these relative abundance values with respect to α-proteobacteria. In addition, genomic dinucleotide relative abundance measures place *Sulfolobus* and *M. capricolum* among the closest to animal mitochondrial genomes, whereas the classical eubacteria, especially the α-proteobacteria, are at excessive distances. There are also considerable molecular and cellular phenotypic analogies among mtDNA, *Sulfolobus*, and *M. capricolum*.

It is widely accepted that the mitochondrial and plastid organelles originated as bacterial endosymbionts (1–3). A central unresolved problem concerns whether mitochondrial evolution is monophyletic or polyphyletic. There is great mitochondrial diversity including extreme size variations and contrasting patterns of mitochondrial genome organization and expression relative to animal, plant, fungal, and protist lineages (1). The current endosymbiont hypothesis, argued largely from rRNA sequence comparisons, proposes that mitochondrial genomes were acquired from a Gram-negative α-proteobacterium with candidate forebears including *Paracoccus denitrificans* (1, 3), *Agrobacterium tumefaciens* (4), or a member of the *Rickettsia* group (2).

Phylogenetic reconstructions from DNA and protein sequences currently determine only the degree of similarity among aligned homologous genes or regions. This is also an indispensable requirement for rRNA gene comparisons. Different evolutionary relationships often result for the same set of organisms from analyses of different gene sequences. We here apply methods of genomic sequence comparisons that do not depend on sequence alignments and that provide assessments of general relatedness of entire genomes.

We will present considerable data supporting the hypothesis that a bacterium of the mycoplasma group, possibly a close relative of *Mycoplasma capricolum* (5), or an archaebacterium like *Sulfolobus solfataricus* or *Sulfolobus acidocaldarius* is a more likely ancestor of the *animal* mitochondrion. Fungal, protist, and plant mitochondrial evolution may have other eubacterial sources. Our methods for assessing genomic similarities are based on analysis of *relative abundance* values of di-, tri-, and tetranucleotides. Genomic sequences are compared with respect to oligonucleotide compositional extremes and dinucleotide relative abundance distances (see *Methods* and Tables 1–3). Further considerations relate to rRNA and tRNA structures, mutation rates and biases, cellular characteristics, special proteins, and energy systems (see Table 4).

METHODS

Data Description. Complete genomes were available for 21 mitochondria and 5 chloroplasts. Sequence sets (most >100 kb) were compiled from a diverse collection of 27 bacterial genomes. Our *Sulfolobus* sequences consist of two closely related species, *S. solfataricus* and *S. acidocaldarius*.

Dinucleotide Relative Abundance Values. A standard assessment of dinucleotide bias is through the odds ratio $\rho_{XY} = f_{XY}/f_X f_Y$, where f_X denotes the frequency of the nucleotide X and f_{XY} denotes the frequency of the dinucleotide XY. The formula for ρ_{XY} is modified for double-stranded DNA by calculating the odds ratio for the given DNA sequence *S* concatenated with its inverted complement sequence (6). In this setting, the frequency f_A of the mononucleotide A in *S* is symmetrized to $f^*_A = f^*_T = (f_A + f_T)/2$ and $f^*_C = f^*_G = (f_C + f_G)/2$. Similarly, $f^*_{GT} = (f_{GT} + f_{AC})/2$, etc. A symmetrized dinucleotide odds ratio measure is $\rho^*_{GT} = \rho^*_{AC} = f^*_{GT}/f^*_G f^*_T$ and similarly for all other dinucleotides. Conservative estimates, $\rho^*_{XY} \geq 1.23$ or ≤ 0.78, indicate when the doublet XY is of significantly high or low relative abundance compared with a random association of its component mononucleotides (7). The corresponding third- and fourth-order measures are $\gamma^*_{XYZ} = (f^*_{XYZ} f^*_X f^*_Y f^*_Z)/(f^*_{XY} f^*_{YZ} f^*_{XNZ})$ and $\tau^*_{XYZW} = (f^*_{XYZW} f^*_{XY} f^*_{XNZ} f^*_{XN_1N_2W} f^*_{YZ} f^*_{YNW} f^*_{ZW})/(f^*_{XYZ} f^*_{XYNW} f^*_{XNZW} f^*_{YZW} f^*_X f^*_Y f^*_Z f^*_W)$, respectively, where N is any nucleotide and W, X, Y, Z are each one of A, C, G, T (7).

Relative Abundance Distances. Consider $\rho^*_{ij} = f^*_{ij}/f^*_i f^*_j$ for all dinucleotide pairs (i, j). We use a measure of dinucleotide "distance" between two sequences f and g, the *dinucleotide relative abundance distance* (δ-distance), calculated as $\delta(f, g) = (1/16)\Sigma_{ij}|\rho^*_{ij}(f) - \rho^*_{ij}(g)|$, where the sum extends over all dinucleotides (7, 8). A third-order trinucleotide relative abundance distance is calculated as $\gamma(f, g) = (1/64)\Sigma_{ijk}|\gamma^*_{ijk}(f) - \gamma^*_{ijk}(g)|$. Corresponding higher order distances are also available (7).

RESULTS AND DISCUSSION

Various genomic compositional properties comparing all available complete mitochondrial sequences with 27 diverse bacterial DNA sets are studied. Table 1 displays di- and tetranucleotide relative abundance extremes for these DNA sets. It is useful first to recall the nature and extent of short oligonucleotide relative abundance extremes in general genomic sequences. For example, the dinucleotide TpA is broadly underrepresented (e.g., refs. 6–9). Apropos, TpA has the least thermodynamic stacking energy (10), entailing flexibility of the TpA site for untwisting the DNA double helix (11). CpG suppression prominent in vertebrate sequences is generally ascribed to the classical methylation/deamination/mutation scenario. The dinucleotide CpG is also distinguished in having the highest thermodynamic stacking energy, possibly suggesting a DNA structural/conformational specificity for CpG (7, 10). The tetranucleotide CTAG is drastically underrepresented in many eubacteria. Interpretations center on structural defects (kinking) associated with this tetranucleotide (6, 7).

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. Extreme relative abundances of some short oligonucleotides in mitochondria and bacteria

Organism*	Size, bp	% G+C	Relative abundance†				
			CpG	CpC/GpG	TpA	GpC	CTAG
			Mitochondria				
Vertebrates							
Human	16,569	44.37	*0.53*†	*1.35*	1.07	0.89	1.10
Cow	16,338	39.39	*0.56*	*1.31*	1.07	0.91	1.10
Whale	16,398	40.59	*0.54*	*1.31*	1.07	0.92	1.00
Seal	16,826	41.72	*0.65*	*1.24*	1.09	0.87	1.05
Rat	16,298	38.68	*0.53*	*1.39*	1.01	0.88	1.08
Mouse	16,295	36.74	*0.52*	*1.36*	1.03	0.94	1.05
Chicken	16,775	45.96	*0.46*	*1.37*	0.99	0.82	1.04
Carp	16,364	43.25	*0.62*	*1.30*	1.05	0.95	1.06
Bonyfish	16,558	45.50	*0.60*	*1.36*	1.05	0.87	1.02
Frog	17,553	36.99	*0.63*	*1.28*	0.99	1.02	1.11
Invertebrates							
A. suum	14,284	28.03	*0.36*	*1.61*	0.83	*0.72*	*1.25*
C. elegans	13,794	23.78	*0.56*	*1.52*	0.97	1.07	1.12
D. yakuba	16,019	21.41	*0.68*	*1.67*	0.95	0.92	*0.56*
Pa. lividus	15,696	39.69	*0.58*	*1.31*	0.93	1.02	1.00
Sy. purpuratus	15,650	41.02	*0.56*	*1.33*	0.92	1.04	0.95
Fungi							
Sa. cerevisiae‡	78,521	17.55	*1.48*	*3.12*	1.22	*1.29*	*1.51*
Sc. pombe	19,431	30.09	*0.54*	*1.32*	0.91	0.94	0.91
P. anserina	100,314	30.06	0.84	*1.25*	1.06	*1.29*	0.90
Protists							
Pm. aurelia	40,469	41.24	0.84	1.17	0.81	*1.20*	0.88
Tr. brucei	23,016	23.30	*0.58*	*1.87*	0.82	1.10	1.02
Plant							
Liverwort	186,608	42.41	0.93	*1.22*	0.85	1.09	0.98
			Chloroplasts				
Plant							
Rice	134,525	38.99	0.86	*1.29*	0.82	0.89	0.92
Tobacco	155,844	37.85	0.87	*1.28*	*0.78*	0.83	0.97
Eu. gracilis	41,017	24.07	1.10	*1.37*	0.85	*1.37*	0.89
Liverwort	121,024	28.81	0.87	*1.38*	0.83	*1.24*	*0.78*
Ep. virginia	70,028	36.00	0.91	*1.43*	0.94	0.92	1.00
			Gram-negative bacteria				
α-Proteobacteria							
Ag. tumefaciens	179,863	52.60	1.18	0.90	*0.66*	1.19	0.87
P. denitrificans	55,242	65.15	1.13	0.89	*0.50*	1.15	*0.20*
R. capsulatus	249,305	65.86	1.19	0.88	*0.33*	1.16	*0.22*
R. sphaeroides	106,312	64.51	1.12	0.90	*0.53*	1.08	*0.42*
Rh. meliloti	258,593	60.17	*1.26*	0.82	*0.53*	1.17	*0.51*
β-Proteobacteria							
N. gonorrhoeae	190,330	51.68	*1.32*	0.99	*0.66*	*1.21*	*0.66*
γ-Proteobacteria							
Az. vinelandii	140,102	64.82	1.10	0.86	*0.48*	1.14	*0.21*
Ha. influenzae	166,617	36.78	1.02	1.01	*0.79*	*1.42*	*0.68*
K. pneumoniae	233,827	57.43	1.17	0.90	*0.79*	*1.29*	*0.34*
Ps. aeruginosa	412,407	62.98	1.09	0.87	*0.59*	1.16	*0.35*
E. coli	1,911,300	51.56	1.17	0.89	*0.74*	*1.26*	*0.25*
Sl. typhimurium	584,624	51.89	*1.24*	0.91	0.82	*1.28*	*0.26*
δ-Proteobacteria							
Mx. xanthus	85,975	67.91	1.05	0.87	*0.44*	1.08	*0.40*
			Gram-positive bacteria				
Ba. stearo.	175,536	49.36	*1.34*	0.95	*0.65*	*1.24*	0.83
Ba. subtilis	1,231,845	43.45	*1.29*	0.81	*0.62*	0.91	0.86
L. lactis	281,299	35.57	0.82	1.03	*0.73*	1.14	0.86
My. leprae	803,847	58.02	1.12	0.88	*0.74*	1.07	0.85
My. tuberculosis	136,978	64.04	1.16	0.89	*0.59*	1.03	*0.77*
St. aureus	328,558	32.61	*1.24*	1.04	0.82	*1.28*	*0.26*
Sr. lividans	101,934	69.87	1.13	0.89	*0.57*	0.97	*0.45*
			Miscellaneous bacteria				
M. capricolum	47,481	29.98	*0.69*	*1.23*	0.86	*1.22*	0.85
Cyanobacterium							
Anabaena sp.	196,614	42.67	0.84	1.05	0.82	1.13	0.94
Spirochete							
B. burgdorferi	126,712	33.23	*0.52*	1.02	*0.76*	*1.36*	0.86
Unassigned							
T. thermophilus	87,995	66.43	*0.74*	*1.24*	*0.68*	0.81	*0.56*
Archaebacteria							
H. halobium	100,572	61.36	*1.29*	0.81	*0.62*	0.91	*0.52*
Me. thermoauto.	66,230	49.50	*0.57*	*1.22*	*0.75*	0.81	*0.41*
Sulfolobus sp.§	106,036	39.22	*0.71*	*1.23*	1.03	0.99	1.01

The Animal Mitochondria–*Mycoplasma* or *Sulfolobus* Connection. What are the arguments for a *Mycoplasma*-like or *Sulfolobus*-like endosymbiont, rather than an α-proteobacterium, giving rise to animal mtDNA? We discuss here compositional extremes and later we analyze relative abundance distances. Focusing on extremes of short oligonucleotide relative abundance values suggests a genomic signature that can relate or discriminate mtDNA with respect to bacterial DNA (Table 2).

(*i*) All animal mitochondria are significantly CpG suppressed, and the same holds for *M. capricolum* and *Sulfolobus* (Table 1). In contrast, virtually all Gram-negative and Gram-positive bacteria display normal or moderately high CpG relative abundances (Table 1). For example, *P. denitrificans* carries CpG modestly on the high side (ρ^*_{CG} = 1.13), patently deviant from the pronounced CpG suppression pervasive in animal mtDNA. This also applies to all other α-proteobacteria examined. However, there are thermophilic bacteria that contain significantly low CpG relative abundances, including the archaebacterium *Me. thermoautotrophicum* (ρ^*_{CG} = 0.57) and the primitive eubacterium *T. thermophilus* (ρ^*_{CG} = 0.75). The spirochete *B. burgdorferi* is also CpG suppressed (ρ^*_{CG} = 0.52). The causes and mechanisms for CpG suppression in animal mtDNA are unknown (12).

(*ii*) Animal mitochondria feature high CpC/GpG relative abundances, and the same holds for *M. capricolum* and *Sulfolobus*. The classical eubacteria are normal in CpC/GpG representations, generally having ρ^*_{CC} somewhat less than 1 (Table 1). Intriguingly, the chloroplast genomes are all significantly high in CpC/GpG relative abundances, which is the only consistent extreme dinucleotide relative abundance of this chloroplast chromosomal collection.

(*iii*) The dinucleotide TpA is broadly underrepresented in most prokaryotic and eukaryotic sequences and markedly low in α- and γ-proteobacteria (Table 1). In contrast, TpA representations are normal across animal mitochondrial genomes and also for *M. capricolum* and *Sulfolobus*.

(*iv*) Relative abundance values for the dinucleotide GpC tend to be on the high side in most α- and γ-proteobacteria but are normal in animal mtDNA and in *Sulfolobus* sequences.

(*v*) The tetranucleotide CTAG relative abundance value is strikingly low in almost all α- and γ-proteobacteria but normal to high in animal mitochondrial genomes and with respect to *M. capricolum* and *Sulfolobus*.

The similarities in the oligonucleotide relative abundance extremes of animal mtDNA with *M. capricolum* and *Sulfolobus*, coupled with the discrepancies of these relative abun-

*Species not listed by their common name are shown with abbreviated genus names. Complete names for these species are *Ascaris suum, Caenorhabditis elegans, Drosophila yakuba, Paracentrotus lividus, Strongylocentrotus purpuratus, Saccharomyces cerevisiae, Schizosaccharomyces pombe, Podospora anserina, Paramecium aurelia, Trypanosoma brucei, Euglena gracilis, Epifagus virginia, Agrobacterium tumefaciens, Paracoccus denitrificans, Rhodobacter capsulatus, Rhodobacter sphaeroides, Rhizobium meliloti, Neisseria gonorrhoeae, Azotobacter vinelandii, Haemophilus influenzae, Klebsiella pneumoniae, Pseudomonas aeruginosa, Escherichia coli, Salmonella typhimurium, Myxococcus xanthus, Bacillus stearothermophilus, Bacillus subtilis, Lactococcus lactis, Mycobacterium leprae, Mycobacterium tuberculosis, Staphylococcus aureus, Streptomyces lividans, Mycoplasma capricolum, Borrelia burgdorferi, Thermus thermophilus, Halobacterium halobium, Methanobacterium thermoautotrophicum.*

†Significance levels ($P \leq 0.001$) for high (≥1.23) and low (≤0.78) compositional extremes are italicized.

‡*Sa. cerevisiae* is anomalous in almost all compositional aspects, mostly due to more than 100 C+G clusters, each about 50–100 bp in length, and large A+T-rich spacers.

§*Sulfolobus* sequences are drawn from *S. solfataricus* and *S. acidocaldarius* in approximately equal proportions.

Table 2. Oligonucleotide relative abundance signatures of mitochondria and various bacteria

Group	% G+C	Relative abundance				
		CpG	CpG/GpG	TpA	GpC	CTAG
Mitochondria						
Vertebrates	−	−−	++	0	0	0
Invertebrates*	−−	−−	++	0	0	0
Fungi†	−−	−	++	0	0	0
Protists	−−	−−, 0	+	−	+	0
Chloroplasts						
All	−	0	++	0	0	0
Bacteria						
Gram-negative α‡	++	0	0	−−	0	−−
Gram-negative γ§	ν	0	0	−−	++, 0	−−
Gram-positive	ν	0, +	0	−−	0, ++	0, −−
Specific bacteria						
M. capricolumn	−−	−−	++	0	+	0
Sulfolobus	−−	−−	++	0	0	0
Me. thermo.	0	−−	+	−−	−	−−
Anabaena	−	0	0	0	0	0
B. burgorf.	−−	−−	0	−−	++	0

C+G content signatures are represented as −−, <40%; −, 40–46%; 0, 47–53%; +, 54–60%, ++, ≥60%. For oligonucleotide relative abundances, signature symbols are denoted as −−, all relative abundances significantly low (≤0.78); −, all relative abundances marginally low (0.79–0.81); 0, all relative abundances in random range (0.82–1.19); +, all relative abundances marginally high (1.20–1.22); ++, all relative abundances significantly high (≥1.23). The symbol ν denotes group variability (low to high). Combinations of symbols reflect differences among the group members. For example, 0, + indicates that most member species are random, while others are marginally high.

***A. suum* is somewhat anomalous with respect to other invertebrates, with significantly low GpC relative abundance and significantly high CTAG relative abundance.

†Excluding *Sa. cerevisiae*. Available mtDNA from *Aspergillus niger* (14,440 bp; % G+C = 26.06) and *Neurospora crassa* (18,323 bp; % G+C = 34.63) were included to increase the number of species in the fungal group. *P. anserina* is the only fungal species with a high relative abundance value for GpC ($\rho^* = 1.29$).

‡*Ag. tumefaciens* differs from the other α-proteobacteria by having average C+G content (53%) and normal representations of CTAG ($\tau^* = 0.87$). *Rh. meliloti* also differs from the other α-proteobacteria with respect to CpG relative abundance ($\rho^* = 1.26$).

§*Sl. typhimurium* differs with respect to CpG and TpA relative abundances ($\rho^* = 1.24$ and 0.82, respectively).

dance extremes relative to all α-proteobacteria, argue against the hypothesis of an α-proteobacterium endosymbiont of animal mitochondria but for the hypothesis of a close relative of *M. capricolum* or *Sulfolobus* as the endosymbiont.

(*vi*) The disparity in overall C+G content between the α-proteobacterium *P. denitrificans* (about 65%) or *Ag. tumefaciens* (about 53%) versus animal mtDNA (21–46%) is large. The mycoplasma and sulfolobus groups are C+G poor to the same extent as mtDNA. Apropos, all α-proteobacteria genomes present a manifest C+G excess (Table 1).

Collectively, the overall C+G content and the relative abundances of dinucleotides and tetranucleotides give each DNA genome a unique signature that is generally constant throughout its genome (8, 13). The factors responsible for this signature are not understood. If (as seems likely) the effect of these compositional properties on the physical chemistry of DNA is the dominant influence, the implication is that each organism and/or its ancestors have experienced different relevant selective inputs. In the absence of strong current selection, the dinucleotide and tetranucleotide compositions should be especially conservative and unlikely to drift with time and, therefore, should frequently serve as good indicators of phylogeny.

Based on the oligonucleotide relative abundance differences, we would postulate a polyphyletic mitochondrial evolution, distinct for plant, protist, and animal mitochondria. The concordance in dinucleotide relative abundance extremes among the animal mtDNA but large variations for fungal mtDNA support the hypothesis that the endosymbiotic origin of animal mtDNA is the most recent such event.

An Animal Mitochondrial Genomic Signature. We propose as a signature for characterizing animal mtDNA several distinctive oligonucleotide relative abundance values. These include measurements of G+C content, the dinucleotide relative abundance values of CpG, CpC/GpG, TpA, and GpC, and the relative abundance value of the palindromic tetranucleotide CTAG. Table 2 displays realizations of the signature for mtDNA and a broad spectrum of bacterial genomes. The animal mitochondria and *Sulfolobus* genomes are in complete accord for the given signature, and *M. capricolum* is substantially in accord. By contrast, mtDNA and eubacteria are highly discordant in these signatures.

Dinucleotide Relative Abundance Distance (δ-Distance) Analysis (See *Methods*). Between-species distances generally exceed within-species distances with concomitant robustness over different parts of the same genome (7, 8, 13). For ease of comparisons, samples of δ-distances are given in the legend of Table 3 between prokaryotic and eukaryotic sequences.

The δ-distances relating animal mitochondrial genomes to sequences of 27 diverse bacterial species finds *Sulfolobus* or *M. capricolum* almost always closest and otherwise the second or third closest (Table 3). Moreover, the determinations of Table 3 place each animal mtDNA farther from α-proteobacteria, generally by a factor of 2 to 3, compared to their δ-distances from *Sulfolobus* and *M. capricolum*. The explicit distances to *P. denitrificans* and to *Ag. tumefaciens* are very large (Table 3), generally more than the distance of human to *Escherichia coli* (see legend to Table 3). The δ-distance of each animal mtDNA to *M. capricolum* and *Sulfolobus* indicates moderate-to-weak similarity. The only exceptions to the striking closeness of the mitochondrion–*M. capricolum*–*Sulfolobus* comparisons are the mitochondria of *Sa. cerevisiae*, which are extreme to all other mitochondria, and *Paramecium*, for which *M. capricolum* and *Sulfolobus* are the second and third closest bacteria.

Dinucleotide Relative Abundance Distances Among Various Bacterial Sequences. The δ-distances among the bacterial sequences place *Anabaena* closest to *M. capricolum*, $\delta = 0.068$ (about the distance of chicken to mouse); next closest are the Gram-positive bacteria *L. lactis* ($\delta = 0.086$) and *St. aureus* (0.088), and equally close is *Sulfolobus* sp. (0.089). The latter are about the distance of human to trout. *M. capricolum* is weakly similar to *B. burgdorferi* (0.101) but is very distant (>0.200) from most Gram-negative bacteria. The closest to the *Sulfolobus* sequences is *M. capricolum* ($\delta = 0.089$) and next closest are the thermophiles *T. thermophilus* (0.106) and *Me. thermoautotrophicum* (0.110). The δ-distances of *Sulfolobus* sp. to all Gram-negative bacteria exceed 0.200. The closest bacterial genomes (distantly related) to *Me. thermoautotrophicum* are *Sulfolobus* ($\delta = 0.110$) and *M. capricolum* ($\delta = 0.114$), and equally close is *L. lactis* ($\delta = 0.115$). Distances to the Gram-negative bacteria are mostly ≥0.200. Unlike the above bacteria, *Me. thermoautotrophicum* is not A+T-rich (Table 1). With respect to δ-distances, *B. subtilis* is closest to *Ag. tumefaciens* ($\delta = 0.056$) and *Ps. aeruginosa* ($\delta = 0.061$).

Relative abundance distance comparisons based on di- and trinucleotides significantly correlate with δ-distances (8). In particular, the closest di- and trinucleotide distances of mtDNA to the bacterial sequences of Table 1 are attained for either *M. capricolum* or *Sulfolobus* (data not shown).

In summary, the genomic δ-distance evaluations overwhelmingly place *M. capricolum*, *Sulfolobus*, and *Me. thermoautotrophicum* singularly close to the animal mitochondrial genomes, whereas the α-proteobacteria are at much greater distances (Table 3).

Table 3. Dinucleotide relative abundance distances between each mitochondrial genome and various bacterial genomes

Mitochondrion	Host*	*M. cap.*	*Sul.*	*Me. ther.*	*Ana.*	*B. bur.*	*L. lac.*	*St. aur.*	*Ba. sub.*	*Ag. tum.*	*P. den.*	α-proteo†	γ-proteo†
Vertebrates													
Human	0.134–0.148	0.106	0.085	0.104	0.141	0.172	0.176	0.152	0.237	0.251	0.281	0.241–0.327	0.201–0.244
Cow	0.129–0.155	0.094	0.070	0.100	0.134	0.165	0.166	0.146	0.223	0.236	0.266	0.226–0.312	0.196–0.229
Whale	NA	0.101	0.079	0.103	0.136	0.167	0.171	0.148	0.222	0.244	0.274	0.234–0.320	0.200–0.237
Chicken	0.124–0.147	0.106	0.081	0.093	0.139	0.156	0.167	0.151	0.228	0.272	0.242	0.239–0.318	0.205–0.235
Mouse	0.140–0.156	0.093	0.080	0.093	0.133	0.165	0.163	0.146	0.209	0.232	0.262	0.223–0.307	0.197–0.225
Rat	0.135–0.178	0.094	0.084	0.088	0.094	0.168	0.162	0.148	0.218	0.231	0.260	0.220–0.306	0.200–0.223
Seal	0.169	0.104	0.095	0.095	0.130	0.188	0.165	0.141	0.222	0.233	0.263	0.232–0.310	0.196–0.226
Bonyfish	0.133	0.082	0.072	0.119	0.130	0.165	0.154	0.143	0.216	0.230	0.266	0.229–0.305	0.177–0.226
Carp	0.138	0.081	0.067	0.102	0.117	0.165	0.153	0.129	0.214	0.228	0.258	0.217–0.303	0.176–0.220
Invertebrates													
X. laevis	0.086–0.124	0.058	0.056	0.093	0.101	0.146	0.130	0.113	0.190	0.204	0.234	0.195–0.280	0.165–0.197
C. elegans	0.213–0.249	0.107	0.119	0.162	0.145	0.188	0.179	0.162	0.250	0.264	0.304	0.266–0.339	0.202–0.263
Ascaris	0.287	0.206	0.266	0.216	0.216	0.241	0.239	0.224	0.284	0.312	0.362	0.363–0.380	0.228–0.333
D. yakuba	0.179–0.193	0.113	0.132	0.191	0.163	0.189	0.180	0.166	0.209	0.233	0.273	0.249–0.312	0.192–0.250
Pa. lividus	0.163	0.078	0.062	0.144	0.127	0.110	0.141	0.152	0.203	0.218	0.272	0.235–0.293	0.193–0.242
Sy. purpuratus	0.163–0.175	0.081	0.076	0.154	0.133	0.111	0.145	0.161	0.202	0.217	0.276	0.243–0.291	0.194–0.251
Fungi													
Sc. pombe	0.124–0.136	0.088	0.054	0.104	0.132	0.131	0.128	0.145	0.197	0.211	0.251	0.213–0.287	0.198–0.226
Sa. cerevisiae‡	0.533–0.539	0.481	0.468	0.534	0.511	0.565	0.532	0.497	0.504	0.485	0.516	0.485–0.534	0.455–0.527
Podospora	0.176	0.084	0.062	0.156	0.122	0.156	0.150	0.128	0.195	0.218	0.253	0.223–0.297	0.152–0.217
Protists													
Trypanosoma	0.174	0.150	0.223	0.217	0.177	0.210	0.192	0.193	0.237	0.246	0.308	0.308–0.311	0.180–0.290
Paramecium	0.267	0.128	0.139	0.219	0.151	0.126	0.151	0.169	0.179	0.214	0.275	0.254–0.263	0.190–0.257
Plant													
Liverwort	NA	0.064	0.096	0.147	0.078	0.132	0.089	0.092	0.119	0.143	0.206	0.171–0.207	0.111–0.177

Formulas for dinucleotide relative abundance distances are given in *Methods*. To provide standards of dinucleotide relative abundance distances, we report several distance evaluations applied to various prokaryotic and eukaryotic sequences. Thus, random sequences (randomly permuted DNA sequences of size about 100 kb) yield mutual distance values about 0.007 within a narrow range. Distances between genomic sequences (samples of 100 kb) from cow relative to genomic sequences of pig average about 0.025, from human to cow about 0.042, from *Sa. cerevisiae* to *Sc. pombe* about 0.036, human to mouse about 0.058, *E. coli* to *Sl. typhimurium* about 0.035, *E. coli* to *Ba. subtilis* about 0.085, human to trout about 0.091, human to *Drosophila melanogaster* about 0.160, human to *E. coli* about 0.211, and *T. thermophilus* to *E. coli* about 0.284. NA, not available.

*Host distance ranges calculated for host samples of size 100 kb. Single values are given when <100 kb of host DNA sequences were available.

†Ranges given for α- and γ-proteobacteria refer to distances between each mitochondrial genome and the bacteria in the respective group. The α- and γ-proteobacteria included in this analysis are listed in Table 1.

‡*Sa. cerevisiae* has an unusual genomic composition, yielding excessively high distance values. See Table 1 for details.

Molecular, Genetic, and Cellular Similarities Among Animal Mitochondria, *M. capricolum*, and *Sulfolobus*. Table 4 itemizes salient phenotypic similarities between animal mtDNA, *M. capricolum*, and *Sulfolobus*. These include the following: (*i*) The low C+G content of mycoplasma and animal mtDNA is often associated with a mutational bias favoring A+T coupled to a reduced genome size. However, there is no trend toward A+T-rich genomes in small viruses of prokaryotic or eukaryotic hosts (19). (*ii*) Codon recognition patterns of *M. capricolum* substantially resemble those of animal mitochondria rather than those of eubacteria (ref. 14, pp. 331–347, 575–591). Moreover, animal mitochondria and *M. capricolum* show pronounced similarities of their tRNA structures (5). (*iii*) The use of UGA in animal and fungal mitochondria to specify the amino acid tryptophan and the mitochondrial codon translation tables are remarkably similar to *M. capricolum* (ref. 14, pp. 575–591). Modification of the universal genetic code tends to isolate the mycoplasmas from horizontal gene exchange. In particular, mycoplasmas do not appear to accept plasmids from other bacteria. Along these lines, change of the genetic code putatively has the effect of preventing complete transfer of the mitochondrial genome into nuclear DNA. (*iv*) It is documented that the mutation rate in vertebrate mtDNA exceeds the nuclear mutation rate by more than 10-fold (1). By contrast, the mutation rate of most plant mitochondrial genomes is substantially smaller than the nuclear DNA mutation rate (1). Along these lines, mycoplasma phylogeny has been characterized by a rapid pace of evolution (ref. 14, pp. 549–559). (*v*) If the mycoplasma–animal mitochondrial connection holds, the capacity of *M. capricolum* to reduce its genome to a minimal genetic system during its evolution putatively affords a capacity of animal mitochondria to further streamline their genomes. (*vi*) A difficulty with the α-proteobacterial endosymbiont origin of mitochondria concerns shedding the bacterial peptidoglycan wall during or after its invasion. It would appear simpler for a wall-less bacterium to penetrate a eukaryotic cell and requisition one or several membrane layers. There is evidence that the vaccinia virus acquires a double membrane coat from the cisternae between the Golgi and the endoplasmic reticulum (15). It seems reasonable that a degenerate genome such as that of the mitochondrial organelle would derive from the smallest, degenerate wall-less bacterium such as *M. capricolum* or that of the small wall-less genome of *Sulfolobus*. (*vii*) The mitochondrial organelle might have been formed as an invaginated compartment containing the invading bacterium. In this context, *Sulfolobus* presents an irregularly lobed cell with the potential to form internal membranes from invaginations of its outer membrane. The resulting structure could resemble the metazoan mitochondrial matrix. (*viii*) *Sulfolobus* appears to have several homologs of the Krebs cycle components typical of animal mitochondria.

Conclusion. Inasmuch as the exclusive use of rRNA genes as molecular chronometers is inadequately justified and sequence comparisons of proteins often do not produce a consistent phylogeny, it is reasonable to employ other measures of relatedness. Among these, comparisons of dinucleotide relative abundances and other genome-wide features analyzed herein correlate well with conventional phylogenies

Table 4. Molecular, cellular, and genome organizational similarities between animal mitochondria, *Mycoplasma*, and *Sulfolobus*

Feature	Animal mitochondria	*Mycoplasma (capricolum/mycoides)*	*Sulfolobus (solfataricus/ acidocaldarius)*
Size	13.8 –17.5 kb; plant mtDNAs are variable and mostly large, 80–2400 kb	600–1200 kb; smallest known genome for free-living organism; considered to have undergone significant reduction in genome size	2250 kb (*S. solf.*), 2760 kb (*S. acid.*); relatively small sizes among bacterial genomes (bottom 10%)
% G+C content	21–46	25–40 (relative to *Mycoplasma* sp.)	30–40
Ancestor (current dogma)	α subdivision of Gram-negative bacteria	Gram-positive progenitor (Lactobacillus group)	Putative direct descendent of primitive bacteria
Phylogenetic classification		Eubacteria; part of diverse group of mycoplasms	Archaebacteria; considered proximal to eukaryote line
Mutation rate	10-fold higher than in nuclear DNA	Thought to be about 2-fold higher than in Gram-positive bacteria	Unknown
Habitat	Endosymbiont, "parasitic"	"Parasitic," different ecological niches, mainly surfaces of animals, insects, and plant tissues	Obligate aerobe; thermoacidophilic, optimal growth at 70–85°C, pH 2
Energy system	Respiration, energy available to host cell	Mainly glycolysis of some sugars: obtain many complex organic molecules from host; carry flavins suggesting some respiration (ref. 14, pp. 181–200)	Primarily sulfur metabolizing; can express several cytochromes (18)
Genes	All genes polycistronically expressed	Most genes constitutive; gene economy	
Special proteins		Reduction in genome size to minimal system; expresses homolog of HSP60 and HSP70	Expresses two kinds of genes: (*i*) eubacteria-like (e.g., glutamine synthetase); (*ii*) eukaryote-like with respect to transcription machinery; encodes a reverse gyrase (17)
Polymerase	1 DNA and 1 RNA pol encoded in nucleus; mt specific	1 or 2 DNA pol 1 RNA pol (eubacteria generally carry 3 DNA pol)	1 RNA pol, similar to eukaryotic pol II; 1 or 2 DNA pol
rRNA	1 operon, 5S unit is lost	1 or 2 operons (classical eubacteria generally have 5–10 operons)	1 operon
tRNA genes	22–24; (tRNA modifications similar to *M. cap.*; e.g., similar anticodon, unmodified A); serine-acceptor tRNA lack D-loop common in tRNA	29 (effectively 26); smallest number of tRNA among bacterial genomes (about 50 tRNAs in classical eubacteria); *Mycoplasma* deficient in modified nucleosides	tRNAs not yet well characterized
Amino acid usages	Arginine usage generally the lowest	Arginine usage very low	Too few genes sequenced for reliable estimates
Codons	All 4-degeneracy families read by single tRNA	All 4-degeneracy families (except threonine) read by single tRNA	
Genetic code	Several differences from universal code, UGA coding for tryptophan	UGA codes for tryptophan	No known alterations in genetic code
Cellular structure	Wall-less (double membrane); steroids adhere to outer membrane of vertebrate mitochondria	Wall-less, extremely plastic, cell morphology attains many shapes	Nearly wall-less, lacks peptidoglycans; irregularly lobed cell shape with flexible membranes; contains or adsorbs steroids (eukaryote-like) in membrane coat

Much of the material in this table concerning mycoplasmas was drawn from several chapters in *Mycoplasmas: Molecular Biology and Pathogenesis* (14). Much of the material concerning *Sulfolobus* was drawn from several chapters in *The Biochemistry of Archaea* (18) and from refs. 16 and 17.

in many cases (7, 8) and appear to be at least as suitable as other measures. This approach leads us to postulate that, among those bacteria presently available for analysis, the closest relatives of animal mitochondria are *M. capricolum* and *Sulfolobus* sp. A number of phenotypic similarities between these bacteria and mitochondria are pointed out. We therefore consider it more likely that the ancestor of the animal mitochondrial genome was related to these two bacteria instead of to other bacterial groups previously proposed. Because all arguments (including this one) are indirect, we do not propose to substitute a new dogma for the current one but only to change the favored working hypothesis.

We acknowledge valuable discussions on the manuscript with Drs. B. E. Blaisdell, P. Bucher, and G. Weinstock. We thank L. R. Cardon for assistance in preparing the tables. This work was supported in part by National Institutes of Health Grants GM51117 (A.M.C.) and HG00335-07 and GM10452-30 (S.K.) and National Science Foundation Grant DMS9403553 (S.K.).

1. Gray, M. W. (1992) *Int. Rev. Cytol.* **141**, 233–357.
2. Gray, M. W. (1993) *Curr. Opin. Genet. Dev.* **3**, 884–890.
3. Dyer, B. D. & Obar, R. A. (1994) *Tracing the History of Eukaryotic Cells* (Columbia Univ. Press, New York).
4. Yang, D., Oyaizu, Y., Oyaizu, H., Olsen, G. J. & Woese, C. R. (1985) *Proc. Natl. Acad. Sci. USA* **82**, 4443–4447.
5. Andachi, Y., Yamao, F., Muto, A. & Osawa, S. (1989) *J. Mol. Biol.* **209**, 37–54.
6. Burge, C., Campbell, A. M. & Karlin, S. (1992) *Proc. Natl. Acad. Sci. USA* **89**, 1358–1362.
7. Karlin, S. & Cardon, L. R. (1994) *Annu. Rev. Microbiol.* **48**, 619–654.
8. Karlin, S. & Ladunga, I. (1994) *Proc. Natl. Acad. Sci. USA* **91**, 12832–12836.
9. Nussinov, R. (1981) *J. Biol. Chem.* **256**, 8458–8462.
10. Breslauer, K. J., Frank, R., Blöcker, H. & Marky, L. A. (1986) *Proc. Natl. Acad. Sci. USA* **83**, 192–196.
11. Travers, A. (1993) *DNA-Protein Interactions* (Chapman & Hall, London).
12. Cardon, L. R., Burge, C., Clayton, D. & Karlin, S. (1994) *Proc. Natl. Acad. Sci. USA* **91**, 3799–3803.
13. Karlin, S., Ladunga, I. & Blaisdell, B. E. (1994) *Proc. Natl. Acad. Sci. USA* **91**, 12837–12841.
14. Maniloff, J., McElhaney, R. N., Finch, L. R. & Basemann, J. B., eds. (1992) *Mycoplasmas: Molecular Biology and Pathogenesis* (Am. Soc. for Microbiol., Washington, DC).
15. Sodeik, B., Doms, R. W., Ericsson, M., Hiller, G., Machamer, C. E., van'tHof, W., vanMeer, G., Moss, B. & Griffiths, G. (1993) *J. Cell Biol.* **121**, 521–541.
16. Heibel, G. E., Anzenbacher, P., Hildebrandt, P. & Schafer, G. (1993) *Biochemistry* **32**, 10878–10884.
17. Benachenhou-Lahfa, N., Forterre, P. & Labedan, B. (1993) *J. Mol. Evol.* **36**, 335–346.
18. Kates, M., Kushner, D. J. & Matheson, A. T., eds. (1993) *The Biochemistry of Archaea* (Elsevier, Amsterdam).
19. Karlin, S., Doerfler, W. & Cardon, L. R. (1994) *J. Virol.* **68**, 2889–2897.