Proc. Natl. Acad. Sci. USA
Vol. 94, pp. 10227–10232, September 1997
Evolution

Compositional differences within and between eukaryotic genomes

SAMUEL KARLIN AND JAN MRÁZEK

Department of Mathematics, Stanford University, Stanford, CA 94305-2125

Contributed by Samuel Karlin, July 9, 1997

ABSTRACT **Eukaryotic genome similarity relationships are inferred using sequence information derived from large aggregates of genomic sequences. Comparisons within and between species sample sequences are based on the profile of dinucleotide relative abundance values (The profile is $\rho^*_{XY} = f^*_{XY}/f^*_X f^*_Y$ for all *XY*, where f^*_X denotes the frequency of the nucleotide *X* and f^*_{XY} denotes the frequency of the dinucleotide *XY*, both computed from the sequence concatenated with its inverted complement). Previous studies with respect to prokaryotes and this study document that profiles of different DNA sequence samples (sample size ≥50 kb) from the same organism are generally much more similar to each other than they are to profiles from other organisms, and that closely related organisms generally have more similar profiles than do distantly related organisms. On this basis we refer to the collection $\{\rho^*_{XY}\}$ as the *genome signature*. This paper identifies ρ^*_{XY} extremes and compares genome signature differences for a diverse range of eukaryotic species. Interpretations on the mechanisms maintaining these profile differences center on genome-wide replication, repair, DNA structures, and context-dependent mutational biases. It is also observed that mitochondrial genome signature differences between species parallel the corresponding nuclear genome signature differences despite large differences between corresponding mitochondrial and nuclear signatures. The genome signature differences also have implications for contrasts between rodents and other mammals, and between monocot and dicot plants, as well as providing evidence for similarities among fungi and the diversity of protists.**

Local and global compositional heterogeneity is recognized on many scales in eukaryotic genomes, including variation in G+C content (e.g., isochore compartments, coding vs. noncoding), mobile insertion elements, characteristic centromeric satellite and telomeric repeats, CpG (=CG) suppression in vertebrates, and methylation patterns. Our recent studies of genomic sequence data have demonstrated that (*i*) the dinucleotide relative abundance values (defined below) of different sequence samples of DNA from the same organism are generally much more similar to each other than they are to sequence samples from different organisms and (*ii*) related organisms generally have more similar dinucleotide relative abundance values than do distantly related organisms (1). Dinucleotide relative abundance values are equivalent to the "general designs" derived from biochemical nearest-neighbor frequency analysis (2, 3). These highly stable DNA doublet forms suggest that there may be genome-wide factors, such as functions of the replication and repair machinery, context-dependent mutation rates, DNA modifications, and base-step conformational tendencies that impose limits on the compositional and structural patterns of a genomic sequence. The set of dinucleotide relative abundance values constitutes a "genomic signature" (1, 4) that may reflect the influence of such factors.

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.

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 ρ^*_{XY} for the given DNA sequence concatenated with its inverted complementary sequence. The set of ρ^*_{XY} values is referred to as the *dinucleotide relative abundance profile*.

A measure of difference between two sequences *f* and *g* (from different organisms or from different regions of the same genome) is the average absolute dinucleotide relative abundance difference calculated as $\delta^*(f,g) = 1/16\,\Sigma_{XY}|\rho^*_{XY}(f) - \rho^*_{XY}(g)|$, where the sum extends over all dinucleotides.

DNA Sequence Samples

Large collections of genomic DNA sequences were extracted from GenBank (Release 101.0, June 1997) for 32 diverse eukaryotic organisms (see legend to Table 1). We restricted attention to species for which at least 100 kb of nonredundant sequence data was available. Most sequence aggregates include several contigs ≥10 kb and often ≥30 kb. Apart from complete chromosomes, the sequences in each species ensemble were culled of duplications. It appears that simple eukaryotes like yeast, nematodes, and *Arabidopsis* consist mostly of closely juxtaposed gene sequences interrupted with rare mobile DNA elements. The current human collection (all contigs ≥80 kb) tends to be biased toward genes of medical interest. The *D. melanogaster* data set is replete with developmental genes. The data collections are certainly unrepresentative of the scope of invertebrate, plant, and protist taxa. For example, the dicots, apart from *A. thaliana*, cover only a few species and the monocots are restricted to a few grasses.

Dinucleotide Compositional Extremes

From statistical theory and data experience, a dinucleotide relative abundance may be conservatively described as significantly low if $\rho^*_{XY} \leq 0.78$ and significantly high if $\rho^*_{XY} \geq 1.23$ (4). We distinguish extremes of dinucleotide relative abundances as follows: extremely high, symbolically +++, $\rho^*_{XY} \geq 1.50$; very high, ++, $1.30 \leq \rho^*_{XY} < 1.50$; significantly high, +, $1.23 \leq \rho^*_{XY} < 1.30$; marginally high, (+), $1.20 \leq \rho^*_{XY} < 1.23$; extremely low, −−−, $\rho^*_{XY} \leq 0.50$; very low, −−, $0.50 < \rho^*_{XY} \leq 0.70$; significantly low, −, $0.70 < \rho^*_{XY} \leq 0.78$; marginally low, (−), $0.78 < \rho^*_{XY} \leq 0.81$.

The following trends were observed (see Table 1).

(*i*) TA is broadly underrepresented in eukaryotes (and prokaryotes) generally in the range $\rho^*_{TA} \approx 0.61$–0.81. The

Abbreviation: mt, mitochondrial.

Table 1. Genome signature extremes for diverse eukaryotes

	G+C	CG	GC	TA	CC GG	TT AA	TG CA	AG CT
Vertebrates:								
homsa	**43%**	**---**		**–**	**+**			
overall ρ*:		0.25		0.74	1.25			
range (67 samples):		0.12–0.45		0.54–0.82	1.17–1.28			
bosta	**40%**	**---**		**–**	**(+)**			
overall ρ*:		0.25		0.74	1.22			
range (6 samples):		0.20–0.29		0.70–0.76	1.18–1.24			
sussc	**48%**	**---**		**––**	**+**			
overall ρ*:		0.33		0.67	1.27			
range (6 samples):		0.25–0.37		0.64–0.72	1.24–1.29			
orycu	**43%**	**---**		**–**	**+**			
overall ρ*:		0.31		0.72	1.23			
range (6 samples):		0.15–0.44		0.64–0.75	1.14–1.27			
musmu	**46%**	**---**		**–**			**+**	**+**
overall ρ*:		0.22		0.72			1.24	1.26
range (21 samples):		0.11–0.28		0.66–0.78			1.17–1.33	1.18–1.35
ratno	**48%**	**---**		**–**			**+**	**+**
overall ρ*:		0.28		0.70			1.23	1.26
range (24 samples):		0.21–0.36		0.67–0.75			1.17–1.27	1.21–1.31
mesau	**50%**	**---**		**––**			**+**	**++**
overall ρ*:		0.29		0.64			1.25	1.30
range (6 samples):		0.28–0.31		0.63–0.65			1.23–1.27	1.28–1.33
galga	**45%**	**---**		**––**			**++**	**+**
overall ρ*:		0.24		0.66			1.31	1.24
range (6 samples):		0.17–0.28		0.61–0.71			1.28–1.36	1.20–1.28
xenla	**41%**	**---**		**(–)**	**(+)**			
overall ρ*:		0.37		0.80	1.21			
range (5 samples):		0.26–0.47		0.79–0.81	1.15–1.25			
Echinoderms:								
strpu	**45%**	**––**		**––**			**(+)**	
overall ρ*:		0.59		0.69			1.20	
range (6 samples):		0.54–0.67		0.61–0.74			1.17–1.25	
Invertebrates:								
drome	**41%**		**+**	**–**		**+**		
overall ρ*:			1.29	0.75		1.23		
range (56 samples):			1.17–1.37	0.69–0.81		1.18–1.32		
drovi	**45%**		**++**	**(–)**				
overall ρ*:			1.46	0.79				
range (6 samples):			1.38–1.52	0.76–0.83				
bommo	**40%**							
caeel	**36%**			**––**		**+**		
overall ρ*:				0.61		1.25		
range (142 samples):				0.54–0.68		1.13–1.36		
Fungi:								
sacce	**38%**	**(–)**		**–**				
overall ρ*:		0.80		0.77				
range (241 samples):		0.74–0.88		0.72–0.84				
klula	**36%**	**–**						
overall ρ*:		0.74						
range (6 samples):		0.71–0.76						
canal	**35%**	**––**		**––**			**+**	
overall ρ*:		0.55		0.69			1.26	
range (6 samples):		0.50–0.61		0.68–0.70			1.24–1.29	
schpo	**36%**			**–**				
overall ρ*:				0.77				
range (32 samples):				0.75–0.80				
neucr	**54%**			**––**				
overall ρ*:				0.61				
range (6 samples):				0.59–0.65				
emeni	**51%**			**–**				
overall ρ*:				0.72				
range (5 samples):				0.70–0.79				
aspni	**52%**			**–**				
overall ρ*:				0.74				
range (6 samples):				0.73–0.74				
ustma	**54%**			**––**	**–**			
overall ρ*:				0.50	0.77			
range (6 samples):				0.47–0.54	0.74–0.79			
Dicots:								
arath	**36%**	**–**		**–**				
overall ρ*:		0.72		0.74				
range (40 samples):		0.60–0.87		0.69–0.80				
nicta	**36%**	**––**		**–**				
overall ρ*:		0.53		0.78				
range (6 samples):		0.49–0.58		0.77–0.80				
soltu	**34%**	**––**		**(–)**				
overall ρ*:		0.58		0.79				
range (6 samples):		0.55–0.62		0.75–0.82				
solly	**33%**	**––**		**(–)**				
overall ρ*:		0.54		0.79				
range (6 samples):		0.52–0.56		0.77–0.82				
Monocots:								
zeama	**45%**	**–**		**(–)**				
overall ρ*:		0.78		0.80				
range (6 samples):		0.76–0.83		0.74–0.84				
horvu	**47%**							
orysa	**42%**	**–**		**–**				
overall ρ*:		0.78		0.78				
range (4 samples):		0.77–0.79		0.76–0.81				
Protists:								
plafa	**25%**	**–**			**++**			
overall ρ*:		0.77			1.38			
range (5 samples):		0.70–0.80			1.24–1.57			
trybr	**46%**			**–**				
overall ρ*:				0.74				
range (4 samples):				0.70–0.81				
dicdi	**28%**	**–**		**––**	**(+)**			
overall ρ*:		0.69		0.71	1.22			
range (6 samples):		0.64–0.73		0.69–0.72	1.16–1.29			

Each sequence collection was cleaned by removing duplicate copies of the same gene as identified in the sequence annotations. Species names are abbreviated as follows: **Vertebrates:** homsa (*Homo sapiens*, human, aggregate sequence length 3,326 kb, minimum contig size 80 kb); bosta (*Bos taurus*, bovine, 136, 10); sussc (*Sus scrofa*, pig, 142, 10); orycu (*Oryctolagus cuniculus*, rabbit, 165, 5); musmu (*Mus musculus*, mouse, 1,065, 50); ratno (*Rattus norvegicus*, rat, 656, 10); mesau (*Mesocricetus auratus*, hamster, 101, 5); galga (*Gallus gallus*, chicken, 297, 10); xenla (*Xenopus laevis*, African clawed frog, 251, 5). **Echinoderms:** strpu (*Strongylocentrotus purpuratus*, sea urchin, 127, 3). **Invertebrates:** drome (*Drosophila melanogaster*, 2,779, 50); drovi (*Drosophila virilis*, 162, 5); bommo (*Bombyx mori*, silkworm, 140, 4); caeel (*Caenorhabditis elegans*, chromosomes II and III). **Fungi:** sacce (*Saccharomyces cerevisiae*, complete genome); klula (*Kluyveromyces lactis*, 117, 3); canal (*Candida albicans*, 165, 3); schpo (*Schizosaccharomyces pombe*, 1,582, 20); neucr (*Neurospora crassa*, 148, 5); emeni (*Emericella nidulans*, 222, 5); aspni (*Aspergillus niger*, 119, 3); ustma (*Ustilago maydis*, 100, 3). **Plants:** arath (*Arabidopsis thaliana*, thale cress, 1,989, 80); nicta (*Nicotiana tabacum*, tobacco, 98, 5); soltu (*Solanum tuberosum*, potato, 138, 3); solly (*Solanum lycopersicum*, tomato, 135, 5); zeama (*Zea mays*, maize, 287, 5); horvu (*Hordeum vulgare*, barley, 114, 5); orysa (*Oryza sativa*, rice, 197, 5). **Protists:** plafa (*Plasmodium falciparum*, malaria parasite, 260, 5); trybr (*Trypanosoma brucei*, 207, 5); dicdi (*Dictyostelium discoideum*, 100, 5). No extremes for the dinucleotides AT, AC/GT, and GA/TC occur in the species at hand. The ranges were determined from ≈50 kb disjoint sequence samples. For sequence collections of aggregate length ≥≈200 kb, the available contigs were concatenated in a random order and the resulting sequence partitioned into equal lengths of ≈50 kb.

lowest average value of ρ^*_{TA} among the eukaryotes is found in the fungus *U. maydis* (0.50). Interestingly, TA occurrences are in the random (normal) range in animal mitochondrial (mt) and chloroplast genomes (5). Possible reasons for TA underrepresentation may be its low thermodynamic stacking energy, which is the lowest among all dinucleotides (6), the high degree of degradation of UA dinucleotides by ribonucleases in mRNA tracts (7), and the presence of TA as part of many regulatory signals (e.g., TATA box, transcription terminators, and polyadenylation signal AATAAA). In this last context, TA suppression may help to avoid inappropriate binding of regulatory factors.

(*ii*) CG is the most underrepresented dinucleotide with drastic suppression in vertebrates on the one hand and significant overrepresentation in some α-proteobacterial and halobacterial sequences on the other hand (8, 9). Overall, ρ^*_{CG} values in vertebrates range from 0.23 to 0.37. A pronounced CG underrepresentation at a similar level is found in the archaeon *Methanococcus jannaschii* (0.32) and also in *Sulfolobus* sequences (8, 9). CG is strongly suppressed in the sea urchin, *S. purpuratus* (0.59), in various yeasts (*S. cerevisiae*, *K. lactis*, and *C. albicans*), and in dicot plants, but is only marginally low to low normal in monocot plants (Table 1). CG is suppressed in animal mitochondria (these ρ^* values are mostly in the range 0.50–0.65), whereas it is in the normal range in plant chloroplast genomes (5). CG has normal representations in insects, worms, and most fungi. CG suppression has usually been ascribed to the classical methylation/deamination/mutation scenario causing mutation of CG to TG/CA. However, this hypothesis cannot account for the pervasive CG suppression in animal mitochondria that lack the standard methylase activity. Moreover, some mammalian genomes and all animal mt genomes have CC/GG high but TG/CA in the normal range suggesting a possible CG → CC/GG mutation bias. We have proposed that CG deficiencies may in some circumstances be due to structural constraints related to high dinucleotide stacking energy, supercoiling, and chromatin packing (1).

(*iii*) The dinucleotides CC/GG, TG/CA, and AG/CT, all a single base mutation from CG, are (except for dicot plants) only overrepresented in genomes with strong CG suppression. Interestingly, scrutiny of Table 1 reveals that these dinucleotide relative abundances separate rodents possessing TG /CA and AG/CT of significantly high representations and CC /GG in the normal range from the nonrodents (primates, artiodactyls, and lagomorphs) that possess relative high abundances of CC/GG, but TG/CA and AG/CT in the normal range.

(*iv*) Other dinucleotide biases in eukaryotes include overrepresentation of GC in *Drosophila* species but apparently not in other higher eukaryotes. GC is significantly abundant in most γ-proteobacteria (8, 9).

(*v*) No dinucleotide extremes were found in the moth *B. mori* or in barley (*H. vulgare*). Protists form a diverse group with no consistent pattern of dinucleotide relative abundances.

Dinucleotide Relative Abundance δ*-Differences Among Eukaryotes

The sequence collections were organized into distinct ≈50 kb sequence samples. The average δ*-differences between samples from the same genome (within-species δ*) or from two different genomes (between-species δ*) are exhibited in Table 2. It is useful to distinguish distinct levels of δ*-differences like *random* ($\delta^* < 0.018$), *very close* ($0.020 < \delta^* < 0.030$), *close* ($0.035 < \delta^* < 0.050$), *moderately similar* ($0.055 < \delta^* < 0.075$), *weakly similar* ($0.080 < \delta^* < 0.115$), *distantly similar* ($0.120 < \delta^* < 0.150$), *distant* ($0.160 < \delta^* < 0.200$), and *very distant* ($\delta^* > 0.200$; cf. ref. 8).

The following relations are evident in Table 2.

(*i*) Within-species δ*-differences are with very few exceptions lower than between-species δ*-differences, reflecting the validity of the genome signature (1, 8, 9). The most homogeneous genomes occur among fungi (see Fig. 1 for *S. cerevisiae*), whereas the most variable genomes are found among protists. The DNA sequence samples of *P. falciparum* are the most heterogeneous (average within-species $\delta^* = 0.059$). Fig. 1 shows distributions of all pairwise δ*-differences between 50 kb sequence samples from several species. The distribution of the δ*-differences between human and mouse sequence samples is only slightly shifted relative to δ*-differences within human sequence samples, reflecting the close relatedness of human and mouse (Fig. 1). On the other hand, the δ*-differences between human and *S. cerevisiae*, and between human and *D. melanogaster* are substantially higher than all within-species δ*-differences.

(*ii*) The δ*-differences were determined for all 50 kb samples between 15 large human contigs each of length at least 80 kb. This collection included Alzheimer disease (STM2) gene (87 kb, on chromosome 1), class II major histocompatibility complex region (198 kb, chromosome 6), T-cell receptor β-chain (685 kb, chromosome 7), tyrosine protein kinase gene (84 kb, chromosome 9), chromosome 12p13 gene cluster (117 kb, chromosome 12), retinoblastoma susceptibility gene (180 kb, chromosome 13), BRCA2 gene region (773 kb, chromosome 13), neurofibromatosis 1 region (101 kb, chromosome 17), breakpoint cluster region (152 kb, chromosome 22), region between the human RCP/GCP and G6PD loci (219 kb, chromosome X), iduronate 2-sulfatase gene region (206 kb, chromosome X), fragile X (FMR-1) gene (153 kb, chromosome X), chromosome X cosmid (106 kb, chromosome X),

Table 2. Average δ*-differences (multiplied by 1,000) within and between eukaryotes based on ≈50 kb sequence samples

hom sa	bos ta	sus sc	ory cu	mus mu	rat no	mes au	gal ga	xen la	str pu	dro me	dro vi	bom mo	cae el	sac ce	klu la	can al	sch po	neu cr	eme ni	asp ni	ust ma	ara th	nic ta	sol tu	sol ly	zea ma	hor vu	ory sa	pla fa	try br	dic di	
67	3	3	3	21	13	2	6	5	3	56	3	3	142	241	2	3	32	3	5	2	2	40	2	3	3	6	2	4	5	4	2	samples
43	37	47	50	56	56	70	69	61	110	173	202	184	169	126	126	111	142	151	153	153	204	124	85	96	86	122	141	121	163	154	126	homsa
	21	40	38	55	52	65	72	55	93	169	198	179	153	109	107	101	131	133	137	133	189	103	71	81	70	114	131	111	157	147	114	bosta
		33	45	66	58	70	82	67	108	174	207	192	157	126	126	104	143	147	158	151	206	117	89	100	87	131	149	126	159	159	114	sussc
			51	67	63	74	77	69	109	168	203	185	155	118	119	115	135	141	145	145	195	114	84	94	83	122	140	119	167	154	122	orycu
				42	40	47	59	77	108	196	223	197	183	138	130	112	157	154	161	153	210	131	94	105	100	134	150	136	180	165	138	musmu
					32	35	59	75	98	197	224	199	177	136	124	107	156	148	154	142	204	125	93	103	98	133	148	135	174	165	132	ratno
						19	58	94	100	209	237	217	172	145	131	115	169	147	154	143	201	132	109	117	110	148	162	149	190	182	146	mesau
							35	88	116	182	204	187	172	133	137	120	142	147	148	156	189	132	102	111	114	121	140	121	202	152	157	galga
								50	89	144	170	143	156	98	94	96	113	127	126	124	178	100	56	66	61	86	104	90	132	124	112	xenla
									51	147	171	138	117	78	62	66	105	77	90	68	132	66	57	56	62	79	87	82	130	107	90	strpu
										35	71	100	99	91	120	127	77	131	122	131	170	122	119	113	117	92	88	81	140	90	112	drome
											31	118	127	121	150	156	102	144	125	155	152	147	147	140	149	112	108	105	178	92	155	drovi
												21	114	79	92	125	73	107	108	105	141	105	109	101	114	78	67	78	125	68	135	bommo
													38	75	94	118	72	68	85	91	117	78	116	109	110	102	91	94	159	92	105	caeel
														21	39	86	36	61	62	56	122	42	54	50	53	41	41	36	113	67	89	sacce
															31	91	62	55	64	44	124	40	54	50	54	49	54	54	108	91	97	klula
																25	104	117	137	113	173	88	60	59	63	99	100	93	128	92	57	canal
																	19	72	67	73	122	59	74	72	75	48	47	42	130	59	100	schpo
																		21	42	41	74	50	85	83	92	68	70	78	148	77	124	neucr
																			31	45	86	64	95	93	96	63	61	70	138	80	140	emeni
																				11	106	48	84	80	89	63	57	72	119	92	122	aspni
																					34	115	142	139	151	116	123	123	207	112	180	ustma
																						30	58	59	63	58	67	59	133	84	92	arath
																							31	27	30	54	67	55	119	86	81	nicta
																								31	32	51	59	51	112	80	77	soltu
																									30	61	69	57	108	96	79	solly
																										32	38	31	111	61	105	zeama
																											31	38	97	58	105	horvu
																												27	112	62	98	orysa
																													59	137	110	plafa
																														34	104	trybr
																															54	dicdi

Non-rodent mammals | Rodents | Bird | Amphibian | Echinoderm | Insects | Worm | Fungi | Dicots | Monocots | Protists

See legend to Table 1 for the list and sizes of the sequence samples.

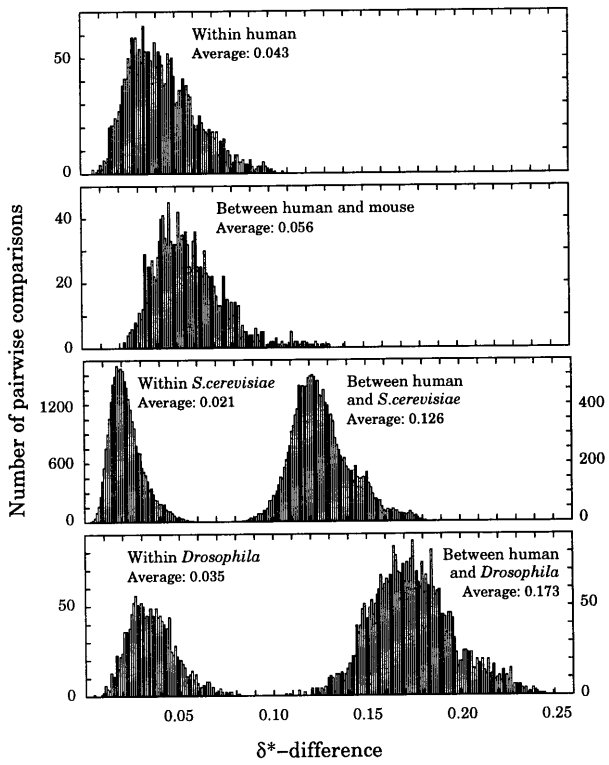

FIG. 1. Distribution of δ*-differences within and between selected species based on all pairwise comparisons of ≈50-kb disjoint sequence samples.

cosmid J136O17 (85 kb, chromosome X), and cosmid J30E17 (82 kb, chromosome X). The average within-contig δ*-differences range from 0.013 to 0.046 and between contigs range from 0.020 to 0.081, which indicates very close to moderate similarity among the human contigs. The highest δ*-differences are revealed between the breakpoint cluster region 152 kb contig of chromosome 22 and other contigs. The ρ^*_{CG} values of these human contigs are drastically low with only one (the 219 kb contig of the X chromosome) above 0.35. This contig is annotated with 17 CpG islands, which may in part explain the marginally higher value $\rho^*_{CG} = 0.41$ versus most values of the other human contigs about 0.25. As with all vertebrates, all contigs are significantly underrepresented in TA at the level ρ^*_{TA} from 0.58 to 0.79 and marginally to significantly overrepresented in CC/GG at the level $\rho^*_{CC/GG} \approx$ 1.20–1.26. The dinucleotide relative abundances of TG/CA and AG/CT are also high normal to significantly high. In ref. 10 we introduced the codon signature, defined as the dinucleotide relative abundances at the distinct codon positions {1,2}, {2,3}, and {3,4} (4 = 1 of the next codon). For large collections of genes (50 or more), we found that the codon signature, like the genome signature, is essentially invariant. Moreover, the codon signature in mammals largely parallels the genome signature but also accommodates amino acid constraints (10).

(*iii*) The available mammals split into two groups: rodents and nonrodents (human, artiodactyls, and rabbit). The mutual δ*-differences within each group are close but δ*-differences between rodents and nonrodent mammals show only moderate similarity, with hamster somewhat more distant from the nonrodents. The chicken and *X. laevis* sequences are weakly or moderately similar to mammals and the sea urchin is weakly similar. In fact, chicken is somewhat closer to rodents than to nonrodent mammals, whereas *X. laevis* is somewhat closer to nonrodent mammals than to rodents.

(*iv*) Insects form a diverse group with mutual δ*-differences in the range 0.071–0.118 (weakly similar). Interestingly, the insects tend to be weakly similar to monocots but generally distantly similar to dicots.

(*v*) Fungi constitute a coherent group with most δ*-differences in the range from 0.035 to 0.075, close or moderately similar. Exceptions are *C. albicans* and the smut *U. maydis*, both distant from other fungi. The δ*-differences between the 16 chromosomes of the yeast *S. cerevisiae* are remarkably close. The within-chromosome δ*-differences range is 0.018 to 0.024 except for chromosome 1 (average δ*, 0.035), and the between-chromosome range is 0.018 to 0.030, indicating that all chromosomes are very close to each other (see Fig. 1). δ*-differences within and between chromosomes II and III of *C. elegans* are remarkably similar with mean δ*-differences about 0.038, marginally higher than the within whole yeast genome.

(*vi*) In the group of plants, the three dicot species (tobacco, potato, and tomato, all in the *Solanaceae* family) are very close (0.027–0.032) and their δ*-differences to the dicot *A. thaliana* are at the level of moderate similarity (0.058–0.063). Our samples from monocots are mutually very close or close and only moderately similar to dicots. Plants (equally with respect to both dicots and monocots) are close to *S. cerevisiae* (δ*-differences 0.036–0.054) and also to the yeast-like *K. lactis*, but only moderately or weakly similar to other fungi. Thus, in terms of DNA normalized doublet comparisons of whole genomes, fungi appear to be closer to plants than to animals, a result that is at variance with some protein sequence comparisons (11). Dicot plants are weakly similar to mammals (δ*-differences 0.070–0.132). On the other hand, insects are distant or very distant to mammals.

(*vii*) Protists are a diverse group. All δ*-differences from *P. falciparum* to other eukaryotes exceed 0.090 and mostly exceed 0.120. *T. brucei* is moderately similar to some insects [the closest δ*(trybr,bommo) = 0.068], some fungi [δ*(trybr,schpo) = 0.059], and monocot plants (δ*-differences 0.058–0.062). *D. discoideum* is weakly similar to yeast and dicot plants.

Comparisons of mt Genomes Between Species

The δ*-differences between various mt-genomes and separately between the nuclear genomes of the corresponding host species are given in Table 3. The principal observation is that levels of similarity among mt genomes assessed by δ*-differences largely parallel levels of evolutionary relatedness assessed by the genome signature δ*-differences among the corresponding host genomes (Fig. 2). That is, species exhibiting small (large) δ*-differences between their nuclear genomes generally have relatively small (large) δ*-differences between corresponding mt genomes and *vice versa*. However, comparing the mt genomic signature with its host genome signature shows δ*-differences generally distant to very distant (δ* values mostly 0.130–0.230, see Table 3) with no discernable pattern among the various eukaryotic kingdoms.

Organellar genomes (mitochondria and chloroplasts) are widely accepted as bacterial endosymbionts in that these genomes constitute the remnants of once free-living cellular organisms (12). There is great diversity among mitochondria, including substantial size variation and contrasting patterns of mt genome organization and gene expression relative to animal, plant, fungal, and protist lineages (12). A central unresolved problem concerns whether mt evolution (primary and/or secondary endosymbiont events) is monophyletic or polyphyletic.

The reported *D. melanogaster* mt genome (19,517 bp) contains about 4 kb (15,566-end) consisting of 12 copies, each about 350 bp long, of the NADH–ubiquinone oxidoreductase chain 6 of unusual composition. This 4 kb section was removed and the remaining sequence used as the adjusted *D. melanogaster* mt genome. The *S. cerevisiae* mtDNA (≈78 kb) composition is an extreme anomaly. This is attested to by the

Table 3. δ*-differences between mitochondrial and the host nuclear genomes (multiplied by 1,000)

homsa	bosta	musmu	ratno	galga	xenla	strpu	drome	caeel	sacce	schpo	arath	trybr	
145	25	26	33	36	50	101	139	85	468	76	158	184	homsa
37	**135**	23	40	48	36	91	130	91	457	68	145	188	bosta
56	*55*	**153**	17	34	37	96	120	87	457	72	154	177	musmu
56	*52*	*40*	**146**	26	44	97	112	88	461	72	154	174	ratno
69	*72*	*59*	*59*	**165**	52	93	130	101	487	68	150	183	galga
61	*55*	*77*	*75*	*88*	**85**	81	106	81	465	66	124	172	xenla
110	*93*	*108*	*98*	*116*	*89*	**172**	116	98	486	52	94	177	strpu
173	*169*	*196*	*197*	*182*	*144*	*147*	**175**	96	385	116	163	167	drome
169	*153*	*183*	*177*	*172*	*156*	*117*	*99*	**228**	439	92	168	131	caeel
126	*109*	*138*	*136*	*133*	*98*	*78*	*91*	*75*	**527**	496	542	388	sacce
142	*131*	*157*	*156*	*142*	*113*	*105*	*77*	*72*	*36*	**130**	84	188	schpo
124	*103*	*131*	*125*	*132*	*100*	*66*	*122*	*78*	*42*	*59*	**83**	240	arath
154	*147*	*165*	*165*	*152*	*124*	*107*	*90*	*92*	*67*	*59*	*84*	**219**	trybr

Included are all eukaryotes for which both the complete mitochondrial genome and a sufficient sample of nuclear genomic DNA is available. δ*-differences between mitochondrial genomes are given in the upper right triangle, δ*-difference between the mitochondrial genome and the corresponding host (boldface) are shown in the diagonal, and average δ*-difference between corresponding nuclear genomes (italics) in the lower left triangle.

exorbitant δ*-differences from all other mt or nuclear genomes (Table 3), in part due to the more than 100 G+C-rich clusters, each about 50 bp long, separated by A+T-rich spacer regions and to the numerous transposable elements. The δ*-differences (Table 3) reflected in the tree (Fig. 2) place *Trypanosoma* far out. It is possible that the mt–*Trypanosoma* endosymbiont reflects a distinct event from that of the other organisms or, like the *S. cerevisiae* endosymbiont, possesses highly anomalous nucleotide organization due to invasion by transposons or other drastic mutational events.

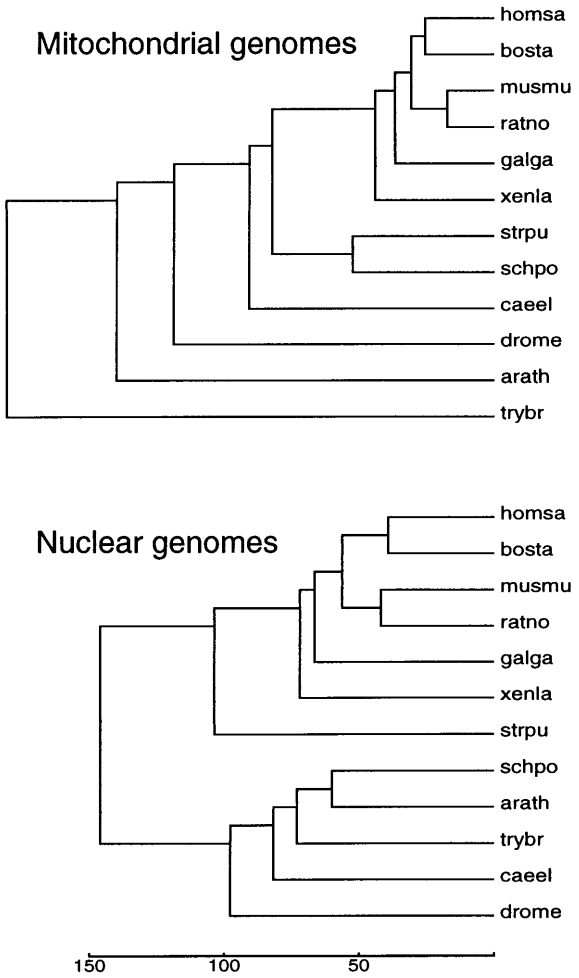

FIG. 2. Evolutionary tree derived from δ*-differences between mt (*Upper*) and nuclear (*Lower*) genomes. The trees were generated by the UPGMA (average linkage) method. The scale (δ*-differences multiplied by 1,000) is shown at the bottom.

Animal (vertebrate and invertebrate) mt sequences show significant underrepresentations of CG dinucleotides, $\rho^*_{CG} \approx$ 0.40 to 0.60 (13), almost to the same extent as occurs in vertebrate genomic sequences. The adjusted *D. melanogaster* and (unadjusted) *D. yakuba* mt genomes entail $\rho^*_{CG} = 0.73$ and 0.68, respectively (13). The fungal *S. pombe* has $\rho^*_{CG} = 0.54$ typical of animal mitochondria. However, the *Podospora anserina* fungal mt CG representation is in the normal range. The mt genome of *A. thaliana* has $\rho^*_{CG} = 0.73$, significantly low. The single persistent significantly high ρ^* value occurs for $\rho^*_{CC/GG} \geq 1.30$ in animal and fungal mt sequences. Intriguingly, the chloroplast genomes are all significantly high in $\rho^*_{CC/GG}$, which is the only consistent extreme among currently available chloroplast sequences (5).

Molecular Evolutionary Implications

Based on comparisons of both the dinucleotide relative abundance extremes (genome signature) and δ*-differences we venture some interpretations of molecular evolutionary relationships among eukaryotic nuclear and mt genomes. In this context, we hypothesize that specificity in replication and repair machinery and context-dependent mutation biases largely maintain the homogeneity of the whole genome of an organism as reflected in the constancy of dinucleotide relative abundances and that differences in this machinery produce the differences in dinucleotide relative abundances among species (1, 9, 14).

(*i*) Rodents are somewhat exceptional among mammals in δ*-differences. In particular, the nonrodent mammals (human, artiodactyls, rabbit) and rodents (mouse, rat, hamster) constitute two coherent groups but show only moderate intergroup similarity. What can account for this separation? Rodents tend to have a higher mutation rate and shorter generation time than many nonrodent mammals (15). Moreover, it is established that rodents are inefficient in global repair of cyclobutane thymidine dimers compared with humans and probably also in repair of other forms of oxidative DNA damage (16). These differences relate principally to replication and repair mechanisms and context-dependent mutation tendencies, consistent with our hypothesis that such molecular differences could produce the observed differences in the genome signature (1, 9). One might inquire about the reasons for differences in repair proficiency of thymidine dimers between human and rodents. We speculate that the rodents analyzed live in more secluded (often underground or noctur-

nal) habitats with lesser exposure to sunlight and other sources of radiation damage. In these environments, we could expect that natural selection has attenuated relevant molecular mechanisms required for efficient repair of cyclobutane dimer adducts. However, embryonic rodent and human cells do not show differences in repair processes and transcription-coupled repair processes are largely similar in humans and rodents (16). The relative abundance value of TT/AA in mouse and rat sequences is about 1.06 (1.03 in hamster) compared with 1.12–1.14 in human, cow, pig, and rabbit. On this basis, there are fewer occurrences of TT (thymidine dimer) in rodents compared with nonrodent mammals.

The foregoing discussion suggests a testable hypothesis. There are rodents active in relatively exposed environments, including the grey squirrel and especially capybara. In these cases, one could expect an efficient cyclobutane dimer repair system. On the other hand, purely nocturnal mammals (e.g., the owl monkey) are conceivably inefficient in thymidine dimer adduct repair.

(*ii*) δ^*-differences in the dicots (tobacco, potato, tomato) are mutually close and similarly for the monocots (maize, barley, rice), whereas the monocot and these dicot sequences are only moderately similar (between-group δ^*-differences larger by a factor of about two than within-group δ^*-differences). Thus, the fundamental monocot/dicot morphological distinction parallels a significant genome signature difference. Intriguingly, the fungi are equally distant to monocots and dicots, whereas the insects are closer to monocots over dicots by a factor of about 1.4. The separation of dicots and monocots (the grasses considered to have arisen from a subgroup of dicots) may have occurred about 200 million years ago (17).

(*iii*) The genome signature comparisons are in agreement with the classic division of most metazoan phyla into the deuterostomes (e.g., vertebrates, echinoderms) and protostomes (insects, worms), since the sea urchin is weakly similar to vertebrates but the protostomes are distant or very distant from the vertebrates.

(*iv*) Compositional biases across bacterial genomes were discussed in Karlin *et al.* (9). The dinucleotide relative abundance values (genomic signature) of bacterial genomes place the *Sulfolobus*-like sequences (eocyte phylogeny) closer to vertebrates than are all other bacterial genomes, and cyanobacteria closer to fungi and plants (4, 8). Enigmatically *Haemophilus* influenzae is moderately similar to *D. melanogaster* sequences.

(*v*) A challenging question concerns reasons and mechanisms to account for the qualitative concordance between the evolutionary development of host nuclear genomes and the development of mt organelle genomes despite the pronounced difference between the mt and host nuclear genome signatures. The mt and nuclear genomes for animal and fungal organisms use independent DNA polymerase machinery (e.g., γ vs. α, ε, δ subunits, respectively). Also, the methods of replication and the nature of the replication origins are fundamentally different. Specifically, the animal and fungal mt transcription-primed replication machinery is distinctive in that most of the heavy strand is synthesized first and the light strand subsequently, whereas the nuclear genomes are replicated analogously to eubacteria synchronized over multiple replication origins.

What about influences of repair processes? There appears to be no DNA excision repair mechanism to deal with cyclobutane dimers in the mitochondrion and apparently bulky lesions are not repaired (18). mtDNA in animals and fungi shows elevated levels of single- and double-strand breaks, mismatches, and generally corrupted base pairings (19). This may be due to a paucity of abasic site correction facilities and mismatch repair capacity in mt genomes (19). Moreover, repair may be less urgent for mt activity because each cell has many mitochondria (hundreds or thousands) and a modicum of impaired organelles may not significantly curtail energy production.

Notably, virtually all mitochondria maintain normal representations of TA dinucleotides, whereas nuclear DNA overwhelmingly tends to have TA in low relative abundance, suggesting that mtDNA may be less thermodynamically stable than nuclear DNA because the dinucleotide TA has the lowest stacking energies compared with all other base steps (6).

We thank Drs. B. E. Blaisdell, A. M. Campbell, and C. Burge for helpful comments on the manuscript. This work was supported in part by National Institutes of Health Grants 2R01GM10452–33 and 5R01HG00335–09, and National Science Foundation Grant 9403553–002.

1. Karlin, S. & Burge, C. (1995) *Trends Genet.* **11,** 283–290.
2. Josse, J., Kaiser, A. D. & Kornberg, A. (1961) *J. Biol. Chem.* **263,** 864–875.
3. Russell, G. J. & Subak-Sharpe, J. H. (1977) *Nature (London)* **266,** 533–535.
4. Karlin, S. & Cardon, L. R. (1994) *Annu. Rev. Microbiol.* **48,** 619–654.
5. Karlin, S. & Campbell, A. M. (1994) *Proc. Natl. Acad. Sci. USA* **91,** 12842–12846.
6. Breslauer, K. J., Frank, R., Blöcker, H. & Marky, L. A. (1986) *Proc. Natl. Acad. Sci. USA* **83,** 3746–3750.
7. Beutler, E., Gelbart, T., Han, J., Koziol, J. A. & Beutler, B. (1989) *Proc. Natl. Acad. Sci. USA* **86,** 192–196.
8. Karlin, S. & Mrázek, J. (1997) in *Bacterial Genomes: Physical Structure and Analysis*, eds. de Bruijn, F. J., Lupski, J. & Weinstock, G. (Chapman & Hall, New York), pp. 196–212.
9. Karlin, S., Mrázek, J. & Campbell, A. M. (1997) *J. Bacteriol.* **179,** 3899–3913.
10. Karlin, S. & Mrázek, J. (1996) *J. Mol. Biol.* **262,** 459–472.
11. Baldauf, S. L. & Palmer, J. D. (1993) *Proc. Natl. Acad. Sci. USA* **90,** 11558–11562.
12. Gray, M. W. (1992) *Int. Rev. Cytol.* **141,** 233–357.
13. Cardon, L., Burge, C., Clayton, D. A. & Karlin, S. (1994) *Proc. Natl. Acad. Sci. USA* **91,** 3799–3803.
14. Blaisdell, B. E., Campbell, A. M. & Karlin, S. (1996) *Proc. Natl. Acad. Sci. USA* **93,** 5854–5859.
15. Li, W.-H., Ellsworth, D. L., Krushkal, J., Chang, B. H.-J. & Hewett-Emmett, D. (1996) *Mol. Phylogenet. Evol.* **5,** 182–187.
16. Bohr, V. A., Phillips, D. H. & Hanawalt, P. C. (1987) *Cancer Res.* **47,** 6426–6436.
17. Savard, L., Li, P., Strauss, S. H., Chase, M. W., Michaud, M. & Bousquet, J. (1994) *Proc. Natl. Acad. Sci. USA* **91,** 5163–5167.
18. Shadel, G. S. & Clayton, D. A. (1997) *Annu. Rev. Biochem.* **66,** 409–434.
19. Yakes, F. M. & Van Houten, B. (1997) *Proc. Natl. Acad. Sci. USA* **94,** 514–519.