Proc. Natl. Acad. Sci. USA
Vol. 91, pp. 12837–12841, December 1994
Evolution

Heterogeneity of genomes: Measures and values

S. KARLIN†, I. LADUNGA‡, AND B. E. BLAISDELL†

†Department of Mathematics, Stanford University, Stanford, CA 94305–2125; and ‡Department of Molecular and Human Genetics, Baylor College of Medicine, Houston, TX 77030

Contributed by S. Karlin, August 24, 1994

ABSTRACT **Genomic homogeneity is investigated for a broad base of DNA sequences in terms of dinucleotide relative abundance distances (abbreviated δ-distances) and of oligonucleotide compositional extremes. It is shown that δ-distances between different genomic sequences in the same species are low, only about 2 or 3 times the distance found in random DNA, and are generally smaller than the between-species δ-distances. Extremes in short oligonucleotides include underrepresentation of TpA and overrepresentation of GpC in most temperate bacteriophage sequences; underrepresentation of CTAG in most eubacterial genomes; underrepresentation of GATC in most bacteriophage; CpG suppression in vertebrates, in all animal mitochondrial genomes, and in many thermophilic bacterial sequences; and overrepresentation of GpG/CpC in all animal mitochondrial sets and chloroplast genomes. Interpretations center on DNA structures (dinucleotide stacking energies, DNA curvature and superhelicity, nucleosome organization), context-dependent mutational events, methylation effects, and processes of replication and repair.**

There are many expressions of genomic heterogeneity: (*i*) local and global variations in C+G content; (*ii*) distinctive direct and inverted repeats, such as REP sequences in *Escherichia coli* (1), telomeric repeats, satellite DNA, and multigene families; (*iii*) transposable elements, such as IS in *E. coli*, Ty in yeast, *Alu* and LINES in human (2); (*iv*) methylation influences (3); (*v*) oligonucleotide relative abundance extremes, such as underrepresentation of the dinucleotide TpA (4, 5) and of the tetranucleotide CTAG in many eubacteria (5, 6); (*vi*) a myriad of control elements (e.g., promoter, enhancer, and termination signals), origins of replication (e.g., automously replicating sequences), and repair recognition sites (e.g., Dam and Dcm in *E. coli*); and (*vii*) genetic mosaicism of genes and genomes resulting from horizontal gene transfer, transposition, and recombination events.

In this paper genomic homogeneity is analyzed and interpreted with respect to short oligonucleotide compositional extremes and dinucleotide relative abundance distances comparing different parts of a genome. The methods are applied to prokaryotic, eukaryotic, and bacteriophage sequences.

METHODS

Data. Current long continuous DNA sequences include a contig of 1.25 Mb centered at *oriC* in *E. coli*, a stretch of 180 kb centered at *oriC* in *Bacillus subtilis*, a 1-Mb stretch of the *Caenorhabditis elegans* genome, and the complete yeast (*Saccharomyces cerevisiae*) chromosome III (YCIII) of 315 kb and chromosome XI (YCXI) of 648 kb. Our analysis concentrates on the following data: the aforementioned chromosomes and contigs, 21 bacteriophage sequences (listed in Table 5), 19 eukaryotic genomic collections mostly exceeding 500-kb aggregate length (Tables 3 and 7), and 21 bacterial DNA sets mostly at least 100 kb long (Table 6). Individual species sequences were combined into aggregations of about 100 kb. A sample sequence is designated *long* when composed from contigs each of length ≥10 kb and designated *short* when composed from contigs of <10 kb. The current human genome collection includes 21 contigs of length 30–180 kb. These *long* contigs were joined, creating 10 *long* samples of lengths 100–125 kb.

Dinucleotide Relative Abundance Values. A common 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} is the frequency of the dinucleotide XY. The formula for ρ_{XY} is modified to accommodate double-stranded DNA by calculating the odds ratio for the given DNA sequence combined with its inverted complement sequence. This changes f_A, the frequency of the mononucleotide A, to $f^*_A = f^*_T = (f_A + f_T)/2$, and similarly $f^*_C = f^*_G = (f_C + f_G)/2$. Also, $f^*_{GT} = (f_{GT} + f_{AC})/2$, etc. The (symmetrized) dinucleotide odds ratio measure for double-stranded DNA is $\rho^*_{AC} = \rho^*_{GT} = f^*_{GT}/f^*_G f^*_T$ and similarly for all dinucleotides. The deviation of ρ^*_{GT} from 1 can be construed as an assessment of dinucleotide bias of GT/AC (7). A corresponding trinucleotide measure 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 for longer oligonucleotides are also available (8). Dinucleotide relative abundances effectively assess contrasts between observed dinucleotide frequencies and those expected from the component mononucleotide frequencies. Similarly, trinucleotide relative abundances appropriately discount the influences of mono- and dinucleotide frequencies, and correspondingly higher-order oligonucleotide relative abundances factor out all lower-order oligonucleotide frequencies.

Dinucleotide Relative Abundance Distance. We use a measure of dinucleotide distance between two sequences *g* and *h* (from different organisms or from different subsets of sequences from the same organism) calculated as $\delta(g,h) = (1/16)\Sigma|\rho^*_{ij}(g) - \rho^*_{ij}(h)|$ (abbreviated δ-distance), where the sum traverses all dinucleotides. The δ-distance contrasts sharply with the *straight dinucleotide frequency distance*, $d(g,h) = (1/16)\Sigma|f^*_{ij}(g) - f^*_{ij}(h)|$, which largely reflects biases in base composition as is apparent in the comparisons of various phage genomes with *E. coli* (Table 1).

Obviously, for the δ-distance the temperate phages (Mu, λ, P1, P4, P22) are among the closest to *E. coli*, whereas the lytic phages (T4, T7, φ29) are more distant (with T7 the farthest). Temperate coexistence apparently produces dinucleotide relative abundance patterns similar to those of the host. For similar analyses and examples dealing with herpesvirus evolution, see ref. 9.

RESULTS

Assessments of Intragenome Homogeneity. For a pair of random DNA sequences, the ρ^*_{XY} values, for any dinucleotide XY, have departure from 1 of the order $1/\sqrt{n}$ for sequences of length *n* (10). Therefore, for $n \approx 100{,}000$, $|\rho^*_{XY} - 1|$ is estimated to be about 0.003, and the δ-distance would average about 0.001. To provide standards, we display in Table 2 δ-distance determinations between several prokaryotic and eukaryotic sequence collections. Table 2 distinguishes distance levels as "very close," "close," "moderately relat-

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. Dinucleotide relative abundance distances (δ) versus dinucleotide frequency distances (d) from *E. coli* to phage

Phage	δ	$d \times 10$	C+G, %
Mu	0.040	0.033	49.4
P4	0.042	0.042	49.5
λ	0.047	0.031	49.8
P1	0.059	0.124	42.8
P22	0.065	0.069	47.4
T4	0.099	0.225	35.7
ϕ29	0.135	0.290	39.7
T7	0.169	0.113	48.4

Consult Table 5 for some information on the phages. The ordering of phage is according to increasing values of the δ-distances from *E. coli*. The C+G content of the *E. coli* genome is about 52%.

ed," . . ., "very distant." The between-species distances are generally greater than the within-species distances (7).

Comparisons of 12 distinct 100-kb sections from a 1.25-Mb contig of *E. coli* give an average δ-distance of 0.016 and range from 0.008 to 0.035 in the category of random to "close." From an aggregate 1165 kb of the *B. subtilis* genome, 21 samples of ≈60-kb length were formed. The average sample δ-distance in this case is 0.033 with range 0.009–0.056. The within-species 100-kb sample δ-distances for *S. cerevisiae* YCIII concatenated with those from YCXI are remarkably small, showing an average value of 0.019 and range of 0.004–0.025. Similarly, *Sch. pombe* δ-distance samples range from 0.007 to 0.023 (Table 3). Within-species distances for the invertebrates *C. elegans*, *D. melanogaster*, and *B. mori* are also persistently small, mostly at the level of "very close" (Table 3). The results for the 100-kb samples obtained by dividing the 1-Mb contig of *C. elegans* into 10 equal sections yield "very close" to "close" mutual δ-distances. The within-species δ-distances for the protist sequences show that *D. discoideum* sample distances average 0.034, somewhat greater than the sample distances within *T. brucei*, average 0.022.

The within-species δ-distances for the vertebrate samples tend to have higher values and a more extended range than those of invertebrate and fungal sequences (Table 3). There are currently in GenBank eight human contigs of length ≥50 kb: major histocompatibility complex (mapped to chromosome 6), 66,109 bp (43.6% C+G); growth hormone and chorionic somatomammotropin genes (chromosome 17), 66,495 bp (49.2% C+G); β-globin region (chromosome 11), 73,326 bp (39.5% C+G); hypoxanthine phosphoribosyltransferase gene (chromosome X), 56,737 bp (40.3% G+C); neurofibromatosis gene (chromosome 17), 100,849 bp; retinoblastoma gene (chromosome 13), 180,388 bp (36.4% C+G); T-cell receptor α gene (chromosome 14), 97,634 bp (44% C+G); and vitamin D-binding-protein gene, 55,136 bp (37% C+G). The sequences exceeding 90 kb were divided into samples of about 50 kb. Dinucleotide relative abundance distances among these human samples produced an average δ-distance of 0.040 and range of 0.011–0.089.

Distances were calculated for samples of about 100 kb from 12 bacterial genomes (Table 4). These are generally in the category of "very close" to "close." Thus, the dinucleotide relative abundance values are closer to random for bacterial, protist, fungal, and invertebrate sequences than for the vertebrate sequences, whose distances may be raised somewhat by mixing isochores of different base composition.

Table 2. Examples of dinucleotide relative abundance distances ($\delta \times 1000$) between genomic collections (see ref. 7)

Comparison	Mean	Range
Random*	7	0–15
Very close (δ = 15–30)		
Bovine vs. pig	23	8–60
Close (δ = 30–45)		
Human vs. bovine	41	10–81
Salmonella typhimurium vs. *E. coli*	39	33–46
Moderately related (δ = 45–65)		
Human vs. mouse	58	24–92
Bombyx mori vs. *Drosophila melanogaster*	59	49–70
Weakly related (δ = 65–95)		
Human vs. trout	92	67–126
E. coli vs. *B. subtilis*	79	74–88
Distantly related (δ = 95–140)		
Trout vs. *D. melanogaster*	123	114–129
Distant (δ = 140–190)		
Human vs. *D. melanogaster*	181	159–227
Very distant ($\delta \geq 190$)		
Human vs. *E. coli*	211	185–263
Thermus thermophilus vs. *E. coli*	265	238–306

Species samples are all ≈100 kb in length. For numbers of samples with each species, see Table 3.
*Three hundred pairs of random sequences of diverse mononucleotide frequencies, each of length about 100 kb, were compared.

Table 3. Dinucleotide relative abundance distances among species samples of various eukaryotic genomic sequences

Organism	No. of samples	$\delta \times 1000$ Mean	$\delta \times 1000$ Range
Vertebrates			
Human	20	35	16–89
Bovine	11	25	8–48
Pig	3	18	11–39
Rabbit	10	28	9–51
Mouse	12	29	9–57
Chicken	14	22	12–45
Xenopus laevis	10	29	10–49
Invertebrates			
D. melanogaster	10	15	6–29
B. mori	2	11	
C. elegans	14	17	9–39
Fungi			
S. cerevisiae	19	14	4–25
Sch. pombe	4	11	7–23
Neurospora crassa	3	11	9–26
Protists			
Trypanosoma brucei	4	22	12–28
Plasmodium falciparum	5	21	12–36
Dictyostelium discoideum	3	34	18–44

Nonredundant species-specific DNA sequence collections were compiled from GenBank. Mitochondrial, rRNA, and tRNA sequences were excluded. Samples of about 100 kb were randomly formed from species-specific sequences. Mean distance and range are calculated over all sample pairs within each species.

Table 4. Mean and range of dinucleotide relative abundance distances among samples (75–125 kb long) within bacterial species

Species	Aggregate length, kb	No. of samples	$\delta \times 1000$ Mean	$\delta \times 1000$ Range
Anabaena	151	2	8	—
Staphylococcus aureus	273	3	16	11–19
Bacillus stearothermophilus	152	2	22	—
Neisseria gonorrhoeae	164	2	15	—
Haemophilus influenzae	144	2	12	—
Salmonella typhimurium	504	6	19	9–34
Klebsiella pneumoniae	148	2	13	—
Agrobacterium tumefaciens	176	2	24	—
Rhizobium meliloti	219	2	30	—
Rhodobacter capsulatus	213	2	10	—
Pseudomonas aeruginosa	345	3	12	8–15
Bacillus subtilis	1165	11	20	7–40

Table 5. Extreme of di- and tetranucleotide relative abundances (× 100) in phage

Type	Phage	Length, bp	TA	AA*	GC	GGGG	CGCG	CCGG	CTAG	GATC	TATA	TCGA	CCGC*	CTAC*	GAAC*	GACC*	GCCC*	Description†
dsDNA	P22	30,002	76		126				33	54	(81)				(122)			Temperate, host *S. typhimurium*
	λ	48,502	71		(120)				52	72								Lysogenic coliphage, cg
	Mu	11,997	69	125	123					63	58	78			(119)			Lysogenic coliphage, transposable
	P1	20,825	(79)		124				39									Temperate, many hosts, episomal
	P4	11,624	73	123	(122)					47					(121)			Temperate coliphage, cg
	T2	8,883				70	74		(81)	26						(122)	123	Lytic coliphage
	T4	97,836					71		(81)	48		77			123	(120)		Lytic coliphage
	ϕ29	19,366				67	11	54		78			142					Lytic, host *B. subtilis*, TBP, replicates linearly, cg
	T3	26,457					70		52	12	64			123	125	132		Lytic coliphage
	T7	39,936				71	75		57	5					127	128		Lytic coliphage, cg
	PRD	14,925		146	154				60	8	59			134	167	147		Lytic, many hosts, contains lipid, replicates linearly, cg
ssDNA fil	F1C	6,407		(120)	(119)	63			26	29	75	00				132		Parasitic coliphage, same as M13, cg
	I22	6,744			133			58	52		73		123					Parasitic coliphage, cg
	IKE	6,883			127	68					73		129				136	Parasitic coliphage, cg
	PF3	5,833	68	123		70	65		63				(119)	125			123	Parasitic coliphage, cg
ssDNA icos	PX1 (ϕ174)	5,386	76	(118)	(129)			65	37	00	67	74		(118)		127		Lytic, host *E. coli* or *S. typhimurium*, cg
	PG4	5,577	(81)					66	29	17	71	76		124		126		Lytic, coliphage, cg
	CP1	4,877				00		37			77		166			(121)	134	cg
dsRNA	ϕ6	13,286	67						43					(122)				Lytic, host *Pseudomonas phaseolicola*,
ssRNA	GA	3,466						50										Lytic coliphage, cg
	MS2	3,569									73							Lytic coliphage, cg

Only those statistically significant relative abundance values that occur in at least two phage for dinucleotides and in at least three phage for tetranucleotides are exhibited. ds, Double-stranded; ss, single-stranded; icos, icosahedral; fil, filamentous. Values in parentheses are marginally significant.
Should be interpreted as the symmetrized oligonucleotide. Thus AA signifies AA/TT; GCGG* signifies GCGG/CCGC, etc.
†cg signifies complete genome available; all phage sequence collections are nonredundant. TBP, terminal binding protein.

Short Oligonucleotide Extremes. Di-, tri-, and tetranucleotide extremes are indicated in Tables 5–7 for bacteriophage, prokaryotic (see also ref. 11), and eukaryotic genomic sequences. We highlight some universals and contrasts.

(*a*) The dinucleotide TpA is broadly underrepresented (4, 5). Among possible reasons are the following. (*i*) TpA has the least thermodynamic stacking energy among all DNA dinucleotides (12). (*ii*) TpA is part of many regulatory sequences (e.g., TATA box, polyadenylation signal) and reduced TpA usage may help to avoid inappropriate binding of regulatory factors. Evidence for untwisting and bending at TpA sites occurs in transcription initiation via protein binding to the TATA box, *Eco*RV binding to its recognition sequence, and γδ resolvase binding at the site at which crossing over occurs (13).

(*b*) The well-known methylation/deamination/mutation scenario can, in part, explain underrepresentation of CpG and overrepresentation of TpG/CpA, certainly in vertebrate sequences (3). In vertebrates, average rates of nucleotide substitutions involving the CpG doublets are among the most rapid (14). The occurrence of the 5-methyl group on cytosine apparently influences the stability and conformation of DNA. In contrast, the GpC pair induces less structural distortions of the helix (15). Replication error rates (including nucleotide misincorporation and transient misalignments) are known to be context dependent (16, 17).

The significant underrepresentation of CpG in three protist genomic sets is intriguing, since the corresponding methylase activity for non-vertebrate eukaryotes has not been detected. CpG is especially low in *Entamoeba histolytica* ($\rho^*_{CG} = 0.35$) and TpG/CpA is significantly high (1.24), paralleling that in vertebrates (Table 7). In contrast, *T. brucei* is comparable in representations to *D. melanogaster* in the normal range (7).

Table 6. Di- and tetranucleotide relative abundance (× 100) extremes in bacterial sequences

Species* (kb†)	TA	AA‡	CG	GC	CTAG	GGCC	ATAG‡	TTAA
Gram-negative bacteria								
A. tum. (175.8)	66							
R. mel. (218.9)	53		126		55		133	
R. cap. (213.3)	34	128			21		152	
N. gon. (164.4)	67	147	131		66			
P. aer. (345.6)	60				38		128	
E. coli (1911.3)	74	(121)		126	26			
H. inf. (144.2)	(79)	(122)		141	69	37		
A. vin. (125.2)	49				23		148	148
M. xan. (64.8)	43				43		173	156
Gram-positive bacteria								
B. sub. (116.0)	60	123		127	78			
B. ste. (151.8)	65	135	133	123				
M. tub. (87.8)	59				72		127	143
S. liv. (272.8)	59				47		152	222
S. aur. (277.8)						65		
Miscellaneous bacteria								
M. cap. (34.4)			73		(80)	78		
B. bur. (78.1)		(120)	68	132		129		
Ana. (151.0)						68		
T. the. (82.9)	69	130	75		56		131	
Archaea								
H. hal. (95.5)	61		130		52			123
M. the. (60.6)	74		58		40		129	
Sul. (97.3)			71			71		

Only extremes present in at least four species are displayed.
Agrobacterium tumefaciens, Rhizobium meliloti, Rhodobacter capsulatus, Neisseria gonorrhoeae, Pseudomonas aeruginosa, Escherichia coli, Haemophilus influenzae, Azotobacter vinelandii, Myxococcus xanthus, Bacillus subtilis, Bacillus stearothermophilus, Mycobacterium tuberculosis, Streptomyces lividans, Staphylococcus aureus, Mycoplasma capricolum, Borrelia burgdorferi, Anabaena sp., *Thermus thermophilus, Halobacterium halobium, Methanobacterium thermoautotrophicum, Sulfolobus* sp.
†Aggregate nonredundant sequence length available.
‡See legend to Table 5.

Table 7. Extreme di-, tri-, and tetranucleotide relative abundances (× 100) in eukaryotic sequences

Species	CA/TG	CC/GG	CG	TA	CCA/TGG	CTAG
D. discoideum		135	72	70	(122)	74
Entamoeba histolytica	124		35	75	124	75
P. falciparum		139	57			
T. brucei				74		
N. crassa				64		
Aspergillus niger				75		72
S. cerevisiae			(80)	77		
Sch. pombe				78		
C. elegans				61		
D. melanogaster				75		79
B. mori						
Trout	123		61	78		
X. laevis	(121)		46	72		
Chicken	123		52	64		
Mouse	123	(120)	26	71		
Pig	(121)		47	60	(121)	
Rabbit	(121)		51	62	123	74
Bovine	(120)		47	65	(121)	
Human		(122)	37	71		

Short oligonucleotides overrepresented (≥120) or underrepresented (≤80) in at least three organisms are listed; extreme values ≥123 or ≤78 are considered significant.

Thermophilic archaea are pervasively CpG suppressed, whereas eubacterial genomes are not and halophilic archaea tend to have overrepresentation of CpG dinucleotides (Table 6). This is consistent with the closer δ-distances observed between archaebacterial thermophiles and vertebrates (8). Unlike eubacteria, GpC relative abundance values for thermophiles are in the normal range. It is striking that all metazoan mitochondrial genomes entail CpG suppression and normal CpA/TpG relative abundance values but significant overrepresentations of CpC/GpG (11). Chloroplast genomes also feature high relative abundance of CpC/GpG (11).

There is a revealing contrast in ρ^*_{CG} and ρ^*_{TA} values for the *long* versus *short* vertebrate samples (Table 8). CpG suppression is more pronounced in the *long* samples (aggregates of longer contigs) than in the *short* samples. Why? The degree of CpG suppression is variable, probably reflecting an irregular distribution of HTF islands (regions of unmethylated CpG), isochore partitions, and biases in nucleosome placements. Most short sequences in GenBank center on genes (coding regions), whereas the *long* contigs contain mostly noncoding regions (introns, spacers, and flanks). In this context, it is likely that the *short* sequences contain more HTF islands and corresponding higher ρ^*_{CG} values than the *long* contigs. Sample sequences composed from *short* contigs derived from many independent sources tend to produce less variable dinucleotide relative abundance profiles (probably a statistical concomitant of the probabilistic law of large numbers). Also, the noncoding parts of *long* sequences in higher eukaryotes often incorporate manifold direct and inverted repeat structures and thus more extremes of relative abundances. The coding parts of *short* sequences may have lower values of TpA because UpA is the RNA dinucleotide most susceptible to RNase activity (26).

Table 8. CpG and TpA suppression in *long* and *short* samples

Sample	Chicken	Mouse	Rabbit	Bovine	Human
ρ^*_{CG} *long*	0.288	0.217	0.499	0.244	0.303
short	0.530	0.358	0.513	0.477	0.413
ρ^*_{TA} *long*	0.653	0.728	0.712	0.774	0.737
short	0.635	0.663	0.602	0.635	0.677

All prokaryotic phages examined carry CpG dinucleotides in the normal relative abundance range. Notably, the temperate phage λ, Mu, P1, P4, and P22, as with enterobacteria (11), exhibit significantly high relative abundances of GpC (Tables 5 and 6).

(*c*) The tetranucleotide CTAG has significantly low relative abundance [τ^*(CTAG) ≤ 0.75] in most eubacteria examined (Table 6). A model of biased DNA repair was proposed (6) to explain the observed low abundance of the CTAG tetranucleotide in the *E. coli* genome. The model attributed the low abundance to the inexact concordance between the specificities of the Dcm methylase and the VSP (very short patch) DNA mismatch repair system. The pervasive rarity of CTAG may implicate a structural role or defect. In this context, the general binding site for the *trp* repressor in *E. coli* contains two copies of CTAG, and there is some evidence from the crystal structure of the *trp* repressor/operator complex that the two CTAGs "kink" when bound by TrpR. Also, the consensus *metJ* repressor binding site involves multiple CTAG sites with kinks (18). It is possible that formation of kinks under supercoiling or other structural conditions is deleterious to DNA stability and that CTAG is therefore avoided or that CTAG may serve some special structural/functional purpose and is therefore used selectively (5). CTAG tetramers cluster in 16S and 23S rRNA genes of many bacteria (19). Moreover, in the rRNA segments the relative positions of the CTAG tetranucleotides are closely conserved. Is it possible that CTAG sites are nucleation or anchor points in the assembly of the ribosomal complex? Interestingly, CTAG tetranucleotides are singularly dense at the lytic replication origin of the human herpesviruses cytomegalovirus and Epstein–Barr virus (data not shown).

(*d*) It is intriguing that the Dam methylase site GATC is of statistically significant low relative abundance in most *E. coli* phage (Table 5). Since all adenine positions of Dam sites of *E. coli* tend to be rapidly methylated after replication, corresponding methylase binding or modification of the DNA would most likely occur at many GATC sites of infecting coliphage, an act presumably hindering phage transcription and replication. On this basis GATC tetranucleotides are presumably selected against in most coliphage.

(*e*) The trinucleotide TAG/CTA is significantly underrepresented in many bacterial (eubacterial and archaeal) genomes (5, 8). In most of these same organisms an excess of TAC/GTA and TGG/CCA relative abundances probably results from avoidance of TAG. Similarly, in eubacteria the overrepresentation of ATAG/CTAT could be a consequence of CTAG avoidance (engendered by the mutation C → A) (Table 6), and the overrepresentation of AATC/GATT in phage T3 and T7 may result from avoidance of GATC (Table 5).

DISCUSSION

Genomic DNA sequences display compositional heterogeneity on many scales, ranging from differences at the isochore level to local signals. In this context, the distribution and stacking of bases along the sequences (e.g., dinucleotide configurations) are of importance. The genome encodes proteins but also must adopt spatial structures propitious for nucleosome placements and chromosomal organization (20, 21). Replication, segregation, transcription, and repair mechanisms are intimately related with structural properties of DNA (15–17, 20, 21). DNA primary structure is not physically or thermodynamically homogeneous, nor is it random (22). Factors that influence DNA oligonucleotide composition and structure include dinucleotide stacking energies, DNA packaging, superhelicity, polymerase nucleotide incorporation biases, and nucleotide modifications.

The influence of the base step (dinucleotide) on DNA conformational preferences is reflected in slide, roll, propeller twist, and helical twist parameters (15, 20). The A, B, and Z

helical forms of DNA appear to depend strongly on base sequences. For example, the base step CC/GG in C+G-rich segments favors an A-form helix, whereas AA/TT exclusively adopts a B form (15). Conformational tendencies for CG and GC steps are for positive slide and negative slide, respectively (15). As a result, poly(CG) can be best accommodated by Z-DNA. Calculations and experiment both indicate that the sugar-phosphate backbones are relatively flexible (15). However, base sequence influences flexural properties of DNA and governs its ability to wrap around histone cores. Moreover, certain base sequences are associated with intrinsic curvature, which can lead to bending and supercoiling (15, 20, 21). Inappropriate juxtaposition or distribution of purine and pyrimidine bases could engender steric clashes (15). For example, transient misalignment during replication is associated with structural alterations of the backbone in alternating purine-pyrimidine sequences. On the other hand, purine and pyrimidine tracts are less reactive, with concomitant reductions in steric conflicts between neighbors (14, 16, 20).

Average dinucleotide relative abundance distances between sequence samples within genomes tend to be small for bacterial, protist, fungal, and invertebrate species and less small for vertebrate species. Average distances between different species genomes (compare Tables 2–4) are generally larger. These inequalities seem to indicate that *some* factors impose limits upon compositional variation of any particular genome. In this context, the 16 (10 symmetrized) dinucleotide relative abundance values provide a robust signature to genomes. Each of the 10 symmetrized dinucleotides exercises its own DNA structural preferences. For example, AA/TT tends to low slide, low role, and high propeller and helical twist; see ref. 15, which describes conformational tendencies of all 10 symmetric dinucleotide types in a B-DNA.

Comparisons were made in terms of di-, di- plus trinucleotide, and di- plus tri- plus tetranucleotide relative abundance distances (7). The di and the corresponding di plus tri (and the di plus tri plus tetra) relative abundance distances between sequences correlate significantly, suggesting that the stacking configurations are principally determined by the base steps, an observation consistent with the thermodynamic stacking energies of short oligonucleotides (12).

Dinucleotide relative abundance deviations putatively reflect duplex curvature, supercoiling, and other higher-order DNA structural features. Many DNA repair enzymes putatively recognize shapes or lesions in DNA secondary structures more than specific sequences (16, 17). DNA structures may be crucial in modulating processes of replication and repair. Nucleosome positioning, interactions with DNA-binding proteins, and ribosomal binding of mRNA are strongly affected by dinucleotide arrangements (20–22).

Other general influences on DNA structure include exposure to sunlight (effects of UV irradiation), osmolarity gradients (e.g., salt concentrations), hydrostatic pressure, acidity and alkalinity tolerance, extreme temperature, alcohol ambience, ecological environment (habitat, energy sources and systems, interacting fauna and flora), and various stress conditions which often trigger transposition events and alternative recombination pathways. There appear to be nucleotide biases in replication, in mutagenesis, and in rates of insertions and deletions dependent on neighboring base context (16, 17). Stacking capacities may influence base incorporation rates and choices. Further factors that impact on genomic structure and organization and flux of DNA involve direct or indirect transfer of genomic pieces between organisms, mediated in part by viruses, bacteria, and animals and by exchanges of plasmids or episomes. Environmental stress appears to enhance conjugation, transposition, transformation, and gene exchange across species lines. It is well known that bacteria can absorb naked DNA. Moreover, *Agrobacterium tumefaciens* transfers plasmids to certain plants, and bacteria can transmit plasmids to yeast through a process like conjugation (23).

DNA has multiple functions, including (*i*) to effect genome replication, (*ii*) to propitiously segregate, and (*iii*) to provide special sequences for encoding gene products. In higher eukaryotes, controls on replication and segregation are not understood and origins of replication may be hardly sequence specific (24). Apropos, there are fundamental differences in replication characteristics between *Drosophila* and mouse (25). *Drosophila* DNA replicates frenetically in the first hour after fertilization, with replication bubbles distributed about every 10 kb. At about 12 hr effective origins are spread to about 40 kb apart. In mouse the rate of replication seems to be uniform throughout developmental and adult stages (24, 25). Moreover, cell divisions involve DNA stacking on itself and loopouts that need to be appropriately decondensed to undergo segregation. The observed narrow limits to intragenomic heterogeneity putatively correlate with conserved features of DNA structure. In this context, the dinucleotide relative abundance signature can discriminate local structure specificity more than sequence specificity.

We express our thanks for discussions on the manuscript from Drs. V. Brendel, C. Burge, A. M. Campbell, and G. Weinstock. This work was supported in part by National Institutes of Health Grants 5R01GM10452-30 and 2R01HG00335-06 and National Science Foundation Grant DMS 91-06974.

1. Lupski, J. R. & Weinstock, G. M. (1992) *J. Bacteriol.* **174,** 4525–4529.
2. Berg, D. E. & Howe, M. M. (1989) *Mobile DNA* (Am. Soc. Microbiol., Washington, DC).
3. Selker, E. U. (1990) *Annu. Rev. Genet.* **24,** 579–613.
4. Nussinov, R. (1987) *J. Theor. Biol.* **125,** 219–235.
5. Burge, C., Campbell, A. M. & Karlin, S. (1992) *Proc. Natl. Acad. Sci. USA* **89,** 1358–1362.
6. Merkl, R., Kroger, M., Rice, R. & Fritz, H.-J. (1992) *Nucleic Acids Res.* **20,** 1657–1662.
7. Karlin, S. & Ladunga, I. (1994) *Proc. Natl. Acad. Sci. USA* **91,** 12832–12836.
8. Karlin, S. & Cardon, L. R. (1994) *Annu. Rev. Microbiol.* **44,** 619–654.
9. Karlin, S., Mocarski, E. S. & Schachtel, G. A. (1994) *J. Virol.* **68,** 1886–1902.
10. Hollander, M. & Wolfe, D. A. (1973) *Nonparametric Statistics Methods* (Wiley, New York).
11. Karlin, S. & Campbell, A. M. (1994) *Proc. Natl. Acad. Sci. USA* **91,** 12842–12846.
12. Breslauer, K. J., Frank, R., Blöcker, H. & Marky, L. A. (1986) *Proc. Natl. Acad. Sci. USA* **83,** 3746–3750.
13. Travers, A. A. & Schwabe, J. W. R. (1993) *Curr. Biol.* **3,** 898–900.
14. Hess, S. T., Blake, J. D. & Blake, R. D. (1994) *J. Mol. Biol.* **236,** 1022–1033.
15. Hunter, C. A. (1993) *J. Mol. Biol.* **230,** 1025–1054.
16. Echols, H. & Goodman, M. F. (1991) *Annu. Rev. Biochem.* **60,** 477–511.
17. Kunkel, T. A. (1992) *Bioessays* **14,** 303–308.
18. Rafferty, J. B., Somers, W. S., Saint-Girons, I. & Philips, S. E. V. (1989) *Nature (London)* **341,** 705–710.
19. Karlin, S., Burge, C. & Campbell, A. M. (1992) *Nucleic Acids Res.* **20,** 1363–1370.
20. Calladine, C. R. & Drew, H. R. (1992) *Understanding DNA* (Academic, San Diego).
21. Wolffe, A. (1992) *Chromatin Structure and Function* (Academic, London).
22. Dickerson, R. E. (1992) *Methods Enzymol.* **211,** 67–111.
23. Heinemann, J. A. & Sprague, G. F., Jr. (1989) *Nature (London)* **340,** 205–209.
24. Krysan, P. J., Smith, J. G. & Calos, M. P. (1993) *Mol. Cell. Biol.* **13,** 2688–2696.
25. Blumenthal, A. B., Kriegstein, H. J. & Hogness, D. S. (1974) *Cold Spring Harbor Symp. Quant. Biol.* **38,** 205–223.
26. Beutler, E., Gelbart, T., Han, J., Kozieol, J. A. & Beutler, B. (1989) *Proc. Natl. Acad. Sci. USA* **86,** 192–196.