30 Hausner, W., Frey, G. and Thomm, M. (1991) *J. Mol. Biol.* 222, 495–508
31 Ouzounis, C. and Sander, C. (1992) *Cell* 71, 189–190
32 Creti, R., Londei, P. and Cammarano, P. (1993) *Nucleic Acids Res.* 21, 2942
33 Frey, G. *et al.* (1990) *Nucleic Acids Res.* 18, 1361–1367
34 Hausner, W. and Thomm, M. (1993) *J. Biol. Chem.* 268, 24047–24052
35 Wettach, J., Gohl, H.P., Tschochner, H. and Thomm, M. (1995) *Proc. Natl Acad. Sci. USA* 92, 472–476
36 Rowlands, T., Baumann, P. and Jackson, S.P. (1994) *Science* 264, 1326–1329
37 Zillig, W. *et al.* (1986) *Syst. Appl. Microbiol.* 8, 197–203
38 Zillig, W. *et al.* (1985) *Nature* 313, 789–791
39 Roberts, S.G.E., Ha, I., Maldonado, E., Reinberg, D. and Green, M.R. (1993) *Nature* 363, 741–744
40 Metz, R. *et al.* (1994) *Mol. Cell. Biol.* 14, 6021–6029
41 Brown, J.W. *et al.* (1988) *Nucleic Acids. Res.* 16, 135–150
42 Thomm, M. and Wich, G. (1988) *Nucleic Acids Res.* 16, 151–157
43 Goodrich, J.A. and Tjian, R. (1994) *Curr. Biol.* 6, 403–409

P. BAUMANN IS IN THE IMPERIAL CANCER RESEARCH FUND CLARE HALL LABORATORIES, HERTS., UK. S.A. QURESHI AND S.P. JACKSON ARE IN THE WELLCOME/CRC INSTITUTE, TENNIS COURT ROAD, CAMBRIDGE, UK CB2 1QR AND IN THE DEPARTMENT OF ZOOLOGY, DOWNING STREET, CAMBRIDGE UNIVERSITY, UK CB2 3EJ.

Dinucleotide relative abundance extremes: a genomic signature

SAMUEL KARLIN AND CHRIS BURGE

Early biochemical experiments established that the set of dinucleotide odds ratios or 'general design' is a remarkably stable property of the DNA of an organism, which is essentially the same in protein-coding DNA, bulk genomic DNA, and in different renaturation rate and density gradient fractions of genomic DNA in many organisms. Analysis of currently available genomic sequence data has extended these earlier results, showing that the general designs of disjoint samples of a genome are substantially more similar to each other than to those of sequences from other organisms and that closely related organisms have similar general designs. From this perspective, the set of dinucleotide odds ratio (relative abundance) values constitute a signature of each DNA genome, which can discriminate between sequences from different organisms. Dinucleotide-odds ratio values appear to reflect not only the chemistry of dinucleotide stacking energies and base-step conformational preferences, but also the species-specific properties of DNA modification, replication and repair mechanisms.

Early biochemical methods of DNA nearest-neighbor (dinucleotide) frequency analysis were applied extensively during the 1960s and 1970s to estimate dinucleotide (base step) frequencies in samples of genomic DNA in many organisms[1–4]. It was observed that the set of dinucleotide odds ratios (dinucleotide frequencies normalized for C+G content, see below) is essentially the same in most organisms for the bulk genomic DNA versus the protein-coding DNA and also essentially the same for DNA fractions of differing sequence complexity (renaturation rate fractions), for euchromatin versus heterochromatin, and for distinct base-compositional (density gradient) fractions of nuclear DNA. These highly stable DNA-doublet odds ratio patterns, referred to by these authors as 'general designs', may in certain respects reflect the total net response of the genome to selection pressures experienced during the evolutionary history of an organism. Recent studies have demonstrated that the general designs of different DNA sequence samples from the same organism are generally much more similar to each other than to sequences from other organisms and that closely related organisms generally have more similar general designs than distantly related organisms[5–7]. These results, together with the earlier biochemical evidence, suggest that there may be factors that impose limits on the compositional and structural variation of a genome, and that the set of dinucleotide odds ratio values constitute a genomic signature.

In this paper, we review the nature of, and mechanisms underlying, the dinucleotide relative abundance (odds ratio) representations, particularly for CG (CpG) and TA (TpA), in a diverse set of prokaryotic, eukaryotic, organelle and viral sequences. The following questions are considered. Are there significant differences in dinucleotide relative abundances between prokaryotes and eukaryotes, between viruses and their hosts, between nuclear and organelle DNA, and between coding, intron and intergenic DNA? How do dinucleotide relative abundance values differ at codon sites (I, II), (II, III) and (III, I)? What about dinucleotides with intervening spaces (XN_kY, where N is any nucleotide and k = 1, 2, 3...)? Do chromatin structure and nucleosome placements impose constraints on base-step associations?

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 in the sequence under study. For double-stranded DNA sequences, a symmetrized version ρ^*_{XY} is computed from frequencies of the sequence concatenated with its inverted complementary sequence[8,9]. Dinucleotide relative abundances ρ^*_{XY} effectively assess contrasts between the observed dinucleotide frequencies and those that are expected from the component mononucleotide frequencies.

TABLE 1. Partial listing of dinucleotide relative abundance values in representative bacterial sequences

Bacteria	Size (bp)	G+C%	CpG	TpA	GpC	ApT	CpC/GpG	ApA/TpT
Gram-negative								
α-Proteobacteria								
R. sphaeroides	106312	64.51	1.12	**0.53**	1.08	**1.38**	0.90	1.04
R. meliloti	258593	60.17	**1.26**	**0.53**	1.17	**1.30**	0.82	1.19
β-Proteobacteria								
N. gonorrhoeae	190330	51.68	**1.32**	**0.66**	(1.21)	0.98	0.99	**1.46**
γ-Proteobacteria								
P. aeruginosa	412407	62.98	1.09	**0.59**	1.16	1.07	0.87	1.16
E. coli	1911300	51.56	1.17	**0.74**	**1.28**	1.10	0.89	(1.21)
Gram-positive								
B. subtilis	1231845	43.45	**1.29**	**0.62**	0.91	1.03	0.81	(1.22)
L. lactis	281299	35.57	0.82	**0.73**	1.14	0.90	1.03	1.18
M. leprae	803847	58.02	1.12	**0.74**	1.07	1.04	0.88	1.04
S. lividans	101934	69.87	1.13	**0.57**	0.97	0.91	0.89	0.90
Spirochetes								
B. burgdorferi	126712	33.23	**0.52**	**0.76**	**1.36**	**0.77**	1.02	(1.22)
Unassigned								
T. aquaticus	42133	63.51	**0.73**	**0.69**	0.85	0.90	**1.26**	**1.29**
Archaebacteria								
H. halobium	100572	61.36	**1.29**	**0.62**	0.91	0.99	0.81	0.96
Sulfolobus spp.	106036	39.22	**0.71**	1.03	0.99	0.96	**1.23**	1.04

Only bacterial genomes with at least 40 kb aggregate sequences were considered. For further details, see Refs 5, 7–9. Relative abundance values meeting conservative thresholds (probability ≤0.001) for high (≥1.23) and low (≤0.78) compositional extremes are in bold. Relative abundances marginally low (0.79–0.82) or marginally high (1.20–1.23) are placed in parentheses. (Ref. Box 1.)

From data simulations and statistical theory, $\rho^*_{XY} \leq 0.78$ (extreme under-representation) or $\rho^*_{XY} \geq 1.23$ (extreme over-representation) occurs for sufficiently long (≥20 kb) random sequences with the probability at most 0.001 for virtually any base composition[8,9]. Several explicit dinucleotide relative abundance values for bacterial sequences are given in Table 1.

CG under-representations

Many classes of organisms exhibit CG under-representation (Tables 1, 2). These include vertebrates, many diverse protist genomes, dicot (but not monocot) plants, metazoan mitochondrial genomes, almost all vertebrate small viral genomes, many examples of thermophilic bacteria and some exceptional bacterial species, e.g. *Borrelia burgdorferi* and *Mycoplasma capricolum*.

CG suppression across vertebrates is commonly ascribed to the classical methylation–deamination–mutation scenario: methylation of CG at position 5 of cytosine, deamination of 5-methylcytosine to thymine and, when unrepaired, conversion to TG/CA. This mechanism might, in part, explain CG deficits and concomitant excesses of TG/CA in vertebrate sequences[10,11]. However, it cannot account for the pervasive CG under-representations in all animal mitochondrial (mt) genomes because invertebrates do not possess the standard methyltransferase and the methylase activity of vertebrate hosts has not been detected in the mitochondrion[12]. Moreover, normal levels of TG/CA are observed in these mt genomes. Of interest is the persistent relative overabundance of the homodinucleotide CC/GG that occurs in animal mt sequences[12,13] and in chloroplast genomes, suggesting a possible CG→CC/GG mutational bias.

All vertebrate small viral genomes (including more than 75 completely sequenced genomes of <30 kb in length) are CG deficient, with the exception of four togaviruses[14]. Examination of these data reveals that, in DNA viruses and retroviruses, the relative abundance of TG/CA is generally in the normal range (i.e. $0.82 \leq \rho^*_{TG} \leq 1.19$), contrary to expectations under the methylation–deamination–mutation mechanism, while CC/GG is often significantly high. Accordingly, simian virus 40, as a free particle, is in an unmethylated state[10], lentivirus genomes are not methylated before integration into the host DNA and murine leukemia virus only becomes methylated weeks after infection[15]. All large or intermediate-size viral genomes (≥30 kb), apart from those of the gammaherpesviruses, have CG relative abundances in the normal range. The gammaherpesviruses [e.g. Epstein–Barr virus (EBV), *Herpesvirus saimiri* and bovine herpesvirus 4] are very CG suppressed and tend to have high TG/CA relative abundances. Various degrees of methylation in different tumorigenic cell lines of EBV have been detected, ranging from unmethylated to an extensively methylated state[6]. By comparison, all sequenced bacteriophages carry CG dinucleotide ratios that are in the normal range. Strikingly, the temperate phages (λ, Mu, P1, P4 and P22) and filamentous parasitic phages (I22, IKe, f_1, f_d and PF1) exhibit significantly high relative abundances of the reverse dinucleotide GC (Refs 5, 9). In addition, γ-proteobacteria of mammalian hosts, but not soil-dwelling γ-proteobacteria, are high in GC (Table 3).

It has been argued that spontaneous mutation rates per nucleotide among living organisms are inversely correlated with genome size[16]. Many DNA viruses have high mutation frequencies and broad adaptability. RNA

TABLE 2. Genomic signature: extremes of dinucleotide relative abundances

	CpG	GpC	TpA	ApT	CpC/ GpG	ApA/ TpT	TpG/ CpA	ApG/ CpT	ApC/ GpT	GpA/ TpC
Prokaryotic organisms										
Gram-negative, α	0, + +	0	− −	+ +, 0	0	0, + +	0	0	− −, 0	v
Gram-negative, β										
(*N. gonorrhoeae*)	+ +	+	− −	0	0	+/+	0	− −	0	0
Gram-negative, γ	0[a]	+ +, 0	− −	0, + +	0	v	0	0	0	0
Gram-negative, δ										
(*M. xanthus*)	0	0	− −	0	0	0	0	0	0	0
Gram-positive	v	0, + +	− −[b]	0	0	0, + +	0	0	0	0, + +
Thermophiles	− −, 0	0	− −, 0	0	+ +, 0	0, + +	0, +	0, − −	0	0
Halobacterium										
(*H. halobium*)	+ +	0	− −	0	0	0	0	0	0	− −
Cyanobacteria	0	0	− −, 0	0	0	0, +	0	0	0	0
Mycoplasma	− −, 0	+, 0	0, −	0, −	+, 0	0, +	0	0	0	0
Spirochetes	− −, 0	0, + +	− −, 0	0, −	0	+ +	0	0	0	0
Phage										
Temperate	0	+ +	− −	0	0	0	0	0	0	0
Lytic	0	0	0, −	0	0	0	0	0	0	0
Parasitic	0	0, +	−, 0	0	0	0, +	0	0	0	0
Organelles										
Mitochondria:										
Metazoa	− −	0, v	0	0	+ +	0	0	0	0	0
Fungi[c]	− −	0, v	0	0	+ +	0	0	0	0	0
Protists	− −, 0	+, v	−	0	+	0	0	0	0	0
Plants	0	0	0	0	0	0	0	0	0	0
Chloroplasts	0	0	0	0	+ +	0	0	0	0	0
Eukaryotic organisms										
Vertebrates	− −	0	− −	0	0, +	0, + +	0, +	0	0	0
Invertebrates	0	0	− −	0	0	0	0	0	0	0
Fungi	0, +	0	− −, 0	− −	0	0	0	0	0	0
Protists	− −, 0	0	− −, 0	0, − −	0	0	0	0	0	0
Plants (grasses)	0	0	− −	0	0	0	0	0	0	0
Plants (dicots)	− −	0	− −	0	0	0	0	0	0	0
Vertebrate										
small viruses ≤30 kb	− −	0	0	0	+ +, 0	0	0, + +	0	0	0
Vertebrate										
large viruses >30 kb	0, − −	0	0, − −	0	0	0	0, +	0	0	0
Plant viruses	− −, 0	0	0	0	0	0	0	0	0	0

The genomic sets of most organisms examined encompass at least 100 kb (many in excess of 500 kb). For each organism, multiple distinct samples of approximately 100 kb were formed and average dinucleotide relative abundance values were determined. For a complete list of the genomes studied, see Box 1.

For dinucleotide relative abundances, the signature symbols are labeled, − −, if all corresponding $\rho^* \leq 0.78$; −, marginally low if $0.79 \leq \rho^* \leq 0.82$; 0, if all corresponding ρ^* fall into the random range 0.83 to 1.19; +, marginally high if $1.20 \leq \rho^* \leq 1.22$; ++, if $\rho^* \geq 1.23$.

Combinations of symbols reflect differences among the group members. For example, 0, + indicates that most member organisms of the group have random ρ^* values and a few are marginally high. The symbol v denotes group variability (low to high).

[a]Exception is *S. typhimurium*, showing $\rho^*_{CG} = 1.24$ and $\rho^*_{TA} = 0.82$.

[b]Exception is *L. lactis*, $\rho^*_{TA} = 0.82$.

[c]Excluding *S. cerevisiae*, available mtDNA from *Aspergillus niger* (14 440 bp) and *Neurospora crassa* (35 546 bp) were included to increase the number of species in the fungal group. *P. anserina* is the only fungal species with high CG ($\rho^*_{CG} = 1.29$). The yeast mitochondrial sequence is anomalous in many compositional respects, mostly due to more than 100 C+G clusters, each about 50–100 bp in length and large A-rich spacers. (Ref. Box 1.)

viruses are substantially error prone, which is generally attributed to the absence of RNA proofreading and mismatch–repair systems. Concomitantly, viral RNA genomes are ubiquitous cellular parasites that tend to replicate quickly and can evolve extremely rapidly. It has been demonstrated experimentally that CG dinucleotides possess the highest thermodynamic stacking energy[17,18]. Reduction in the CG frequencies (thus reduced stacking energy) might facilitate DNA replication and transcription, which would be advantageous for organisms of small genome size needing to replicate rapidly. The high relative abundance of CC/GG in most RNA viral genomes could, therefore, be a result of selection against CG (Ref. 14).

TABLE 3. Dinucleotide relative abundance (ρ_{XY}) values in human coding, intron and intergenic DNA

XY	Coding (I, II)	Coding (II, III)	Coding (III, I)	Intron	Intergenic
AA	1.16	1.16	0.94	1.14	1.16
AC	0.87	**0.77**	0.79	0.80	0.80
AG	0.97	**1.24**	**1.31**	1.22	1.19
AT	0.95	0.88	0.82	0.85	0.85
CA	0.89	**1.36**	**1.30**	1.19	1.15
CC	1.04	1.21	1.11	**1.30**	**1.27**
CG	**0.76**	**0.39**	**0.47**	**0.27**	**0.36**
CT	**1.28**	**1.34**	**1.39**	1.21	1.19
GA	1.17	0.98	0.94	0.99	0.98
GC	0.96	1.17	1.07	0.96	0.98
GG	1.16	0.88	1.13	**1.26**	**1.26**
GT	**0.72**	0.90	**0.73**	0.82	0.79
TA	**0.57**	**0.51**	**0.57**	**0.73**	**0.72**
TC	**1.23**	0.94	0.85	0.97	0.97
TG	1.11	**1.35**	**1.53**	1.18	1.17
TT	1.21	0.90	0.87	1.12	1.16

For columns labeled 'Coding (I, II) (II, III) or (III, I)', a collection of 386 nonredundant human complete gene sequences (total of 130 515 codons) was collected and analyzed as follows. Frequencies of each mono- and dinucleotide were tabulated in each of the three codon positions (designated as $f_X^{(1)}$, $f_{XY}^{(3,1)}$, etc.). Odds ratios were calculated as, for example: $\rho_{CT}^{(2,3)} = f_{CT}^{(2,3)}/f_C^{(2)}f_T^{(3)}$. For the last two columns of the table, collections of nonredundant human introns (1345 intron sequences totaling 1 353 909 bp) and of nonredundant human flanking and intergenic sequences (481 sequences totaling 771 797 bp) were used. Values of $\rho \leq 0.78$ or ≥ 1.23 are printed in bold (see legend to Table 1).

Thermophilic Archaea and other thermophilic bacteria (Figs 1, 2) are mostly CG suppressed, consistent with the dinucleotide relative abundance DNA similarities observed between thermophiles and vertebrates[9]. The under-representation of CG in various protist genomic sequences (*Dictyostelium discoideum*, ρ^*_{CG} = 0.72; *Entamoeba histolytica*, 0.35; *Plasmodium falciparum*, 0.57; *Tetrahymena* spp., 0.78) is intriguing because the CG methylase activity is not present in invertebrates and the TG/CA frequencies generally are not high. The protist genomes of *Giardia lambdia*, *Toxoplasma gondii*, *Acanthamoeba* spp. and *Trichomonas* spp. are not CG suppressed. These results emphasize the extreme diversity among protists with respect to the genomic organization and composition.

Examination of DNA sequences from plant species (four monocot grasses, seven diverse dicots and two algae; Table 3), reveals that plants are persistently under-represented in TA ($0.47 \leq \rho^*_{TA} \leq 0.78$). For CG relative abundances, there is a contrast between dicots and monocots: the dicots are markedly CG suppressed, $0.50 \leq \rho^*_{CG} \leq 0.76$, but the monocot grasses carry low normal frequencies ($0.84 \leq \rho^*_{CG} \leq 0.88$). The methylation activity in flowering plants acts on cytosine in the context CG or C*N*G and occasionally at C alone[19]. Methylation frequency in maize is high, estimated at more than 25% of all C bases. The two algae *Chlamydomonas reinhardtii* and *Euglena gracilis* are both very low in TA, but only *E. gracilis* is CG suppressed (ρ^*_{CG} = 0.71, compared with ρ^*_{CG} = 0.89 in *C. reinhardtii*). Both algae are significantly high in TG/CA ($\rho^*_{TG/CA}$ = 1.26 or 1.27, respectively).

In many cases of CG suppression, the highest dinucleotide relative abundance is achieved for CC/GG (often significantly high) and not for TG/CA. This inequality applies even where the standard methylase is active. For example, consider the three largest available human genomic contigs: for the region of the retinoblastoma gene (HUMRETBLAS, 180 kb), ρ^*_{CG} = 0.24, $\rho^*_{CC/GG}$ = 1.25, $\rho^*_{TG/CA}$ = 1.16; for the region of the fragile X mental retardation gene (HUMFMR1S, 152 kb), ρ^*_{CG} = 0.23, $\rho^*_{CC/GG}$ = 1.23, $\rho^*_{TG/CA}$ = 1.19; for the T-cell receptor β gene (HUMTCRB, 684 kb), ρ^*_{CG} = 0.18, $\rho^*_{CC/GG}$ = 1.22, $\rho^*_{TG/CA}$ = 1.23. The degree of CG suppression in vertebrates is variable, reflecting an irregular distribution of *Hpa*II tiny fragment islands (HTF; regions of unmethylated CG) and isochore compartments.

TA under-representations

The dinucleotide TA is under-represented across prokaryotes and eukaryotes (Tables 2, 3). Exceptions include most metazoan, fungal and plant mitochondrial genomes and chloroplast genomes. Gram-negative α-proteobacteria are low in TA dinucleotides with ρ^*_{TA} values in the range 0.30–0.66. Eukaryotic genomes are also TA suppressed (ρ^*_{TA} values in the range 0.60–0.80). The reverse dinucleotide, AT, has normal representations in most organisms (i.e. ρ^*_{AT} ~1), except in the α-proteobacteria, where AT tends to be over-represented.

The under-representation of TA is almost universal and might be related to the stacking energy of TA, which is the lowest of any dinucleotide[17,18], which would provide flexibility for unwinding of the DNA double helix. Evidence for untwisting and bending at TA sites occurs in transcription initiation via protein binding, e.g. binding of TFIID at the TATA box, binding of *Eco*RV to its recognition sequence, GATATC, and binding of the γδ resolvase at cross-over points[20]. TA is part of many regulatory sequences (e.g. TATA box, polyadenylation signals: AATAAA in higher eukaryotes, TATATA in yeast), so restricted TA usage may help to avoid inappropriate binding of regulatory factors.

Coding versus noncoding CG and TA representations

Table 3 reports all single-strand dinucleotide relative abundance values in human coding sequences covering 130 515 nonredundant codons. (The results were further corroborated for ten distinct collections, each exceeding 100 000 codons.) The ρ_{TA} values are uniformly low (0.54, 0.56 and 0.61) in codon positions (I, II), (II, III) and (III, I), respectively. In human introns and intergenic regions, TA suppression is maintained at the somewhat higher level $\rho_{TA} \approx 0.72$.

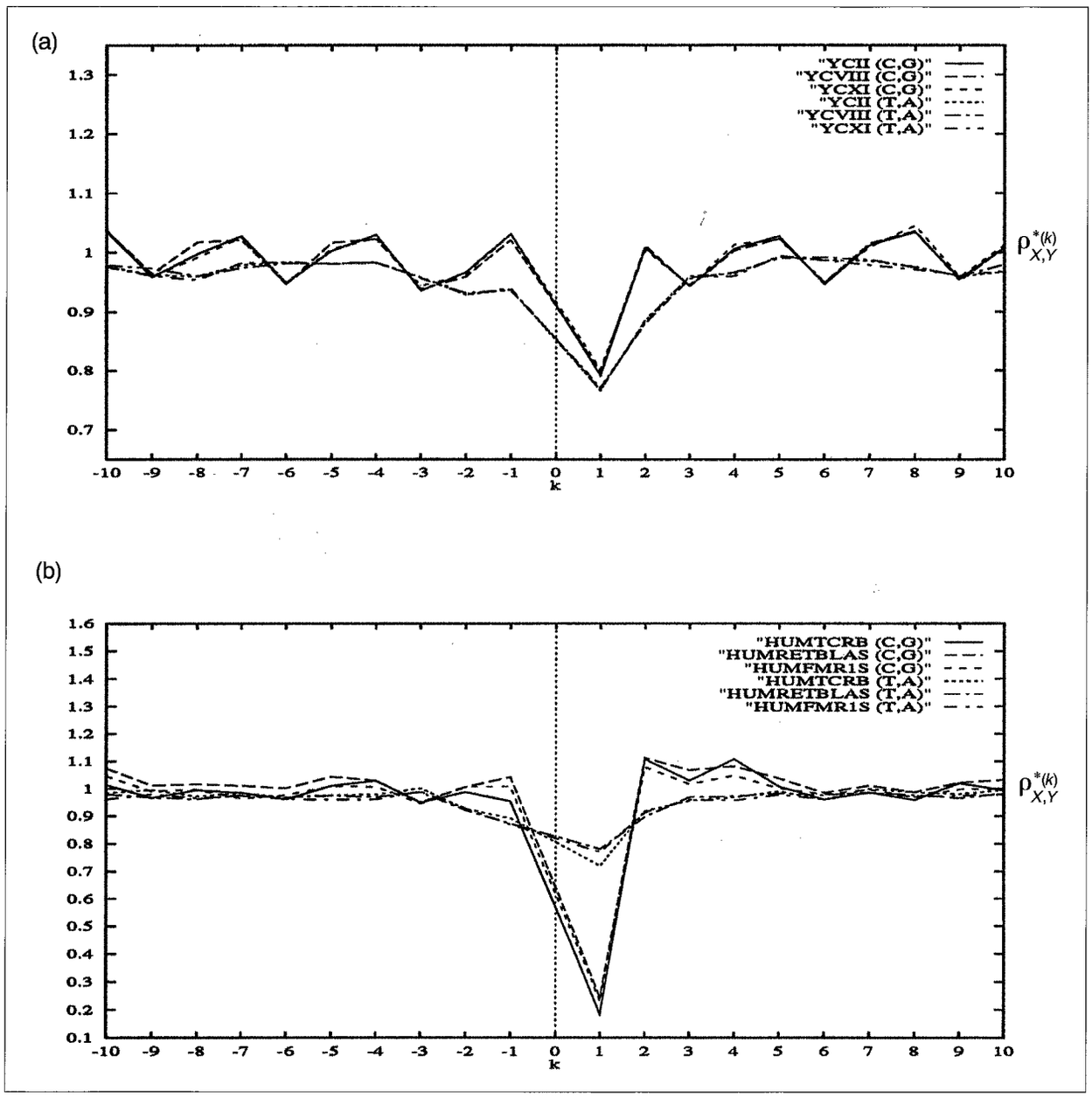

FIGURE 1. Shows a plot of $\rho^{*(k)}_{X,Y}$ for $(X,Y) = (C,G)$ and $(X,Y) = (T,A)$ in (a) the yeast chromosomes II (807 kb), VIII (562 kb) and XI (666 kb) and (b) the human GenBank files HUMTCRB (human T-cell β receptor gene region; 684 kb), HUMRETBLAS (human retinoblastoma gene region; 180 kb) and HUMFMRIS (human fragile X mental-retardation gene region; 152 kb). In (a) and (b), the values of $\rho^{*(k)}_{(C,G)}$ and $\rho^{*(k)}_{(T,A)}$ are plotted for $k = \pm1, \pm2, \ldots \pm10$, where $\rho^{*(k)}_{(X,Y)} = \rho^{*}_{XN_{(k-1)}Y}$ for $k \geq 1$ and $\rho^{*(k)}_{(X,Y)} = \rho^{*}_{YN_{(-k-1)}X}$ for $k \leq -1$.

Accordingly, Beutler *et al.*[21] have shown that UpA is the RNA dinucleotide that is most susceptible to RNase activity. The ρ_{CG} values at the codon positions (I, II), (II, III) and (III, I) are 0.70, 0.36 and 0.42, respectively, and in introns $\rho_{CG} \approx 0.27$. The higher value at codon sites (I, II) undoubtedly reflects requirements of arginine usage. In human proteins, arginine usage derived from CGN codons occurs, on average, at a frequency of 3.2% and arginine encoded from AGR codons occurs, on average, at a frequency of 2.2%. The very reduced ρ_{CG} level in introns and intergenic sequences (except in the HTF islands) presumably relates to the higher substitution rates in the noncoding genomic regions.

In metazoan mitochondrial genomes, CG is consistently under-represented at codon sites (I, II), (II, III) and (III, I), with the lowest values at codon sites (I, II) (Ref. 12). This reflects an extremely low arginine usage in these mitochondria. The retroviruses are uniformly CG suppressed, independent of codon sites and of coding, or noncoding, regions[14].

Dinucleotide relative abundance values with spacings

How do XN_kY (k = 0, 1,...) dinucleotide relative abundance values compare for different spacings (k)? The variation in $\rho^{*}_{TN_kA}$ and $\rho^{*}_{CN_kG}$ are displayed graphically (Fig. 1) for three *Saccharomyces cerevisiae* chromosomes and three long human contigs. The yeast

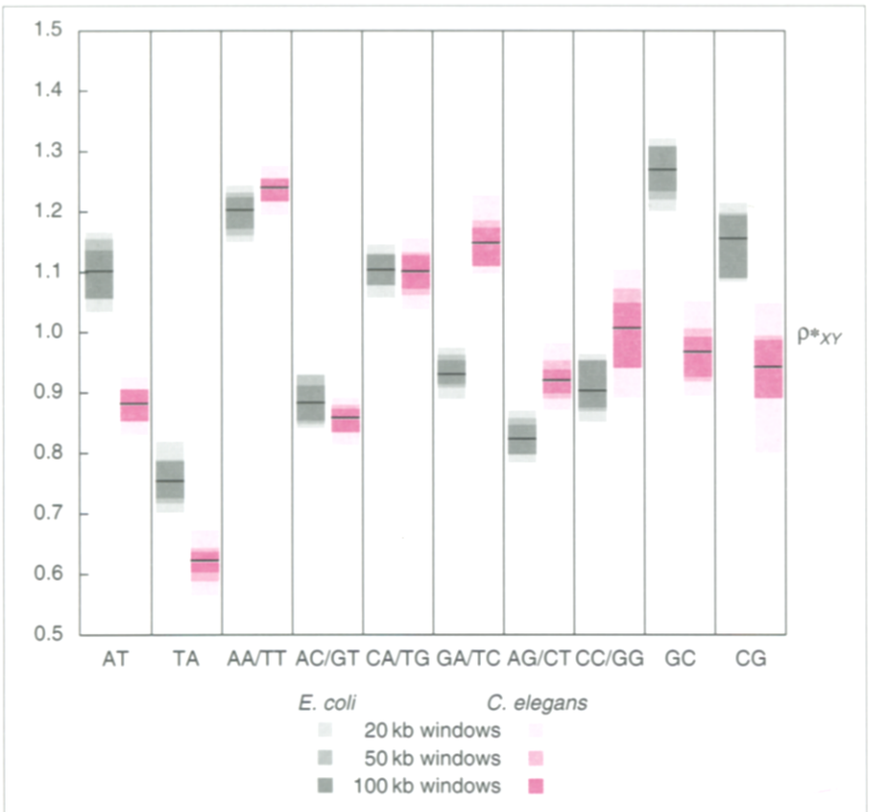

FIGURE 2. The 'general designs' of *Escherichia coli* and *Caenorhabditis elegans*. A 1.25 Mb contig from *E. coli* and a 1.06 Mb contig from *C. elegans* were partitioned into disjoint sequence windows of 20 kb, 50 kb and 100 kb. For each set of sequence windows, the mean and 90% range (5th percentile to 95th percentile) of ρ^*_{XY} was calculated for each dinucleotide *XY*. The mean values are shown as horizontal black lines, the 90% ranges as vertical gray or purple bands. The three window sizes are shown as different shades of gray (*E. coli*) or purple (*C. elegans*); the darker bands corresponding to larger window sizes are superimposed on the lighter bands. For each dinucleotide, the set of bands for the *E. coli* contig (gray) is shown to the left of the set of bands for the *C. elegans* contig (purple).

chromosomes and three long human contigs. The yeast chromosomes II, VIII and XI have ρ^*_{TA} = 0.78, 0.79 and 0.77, respectively, and no other significant extremes. The dinucleotide relative abundance values of YCII, YCVIII and YCXI are remarkably close for all XN_kY. The $\rho^*_{TN_kA}$ and $\rho^*_{CN_kG}$ curves of the three large human contigs are also strikingly similar. A similar congruence applies generally in comparing any two large contigs from the same genome (data to be presented elsewhere). These results suggest that the most significant dinucleotide departures from randomness are observed for k = 0 and that the genomic base step (dinucleotides of zero gap) embodies the most important determinants of DNA structure–function.

DNA structures and dinucleotide representations

The dinucleotide relative abundance deviations observed in Tables 1 and 2 putatively reflect base-step stacking capacities, duplex curvature and other higher-order DNA structural features. Theoretical calculations[22] present methods for predicting dinucleotide sequence-dependent DNA structure. These take account of the energetics of base-step stacking, cross-strand steric clashes (e.g. at pyrimidine–purine steps), and electrostatic interactions determined by the distribution of partial atomic charges within the base pairs and the π-electrons of their aromatic rings (e.g. GG/CC charge repulsion)[20,22,23]. Base-step conformational preferences are reflected in slide, roll, tilt, propeller twist and helical-twist parameters. For example, CG and GC steps carry strong preference for positive and negative slide, respectively[22,23]. The base step CC/GG in C+G rich segments tends to exhibit large negative slide, slight positive roll and variable twist, favoring an A-form DNA conformation, whereas AA/TT steps primarily adopt a B-form[22,24].

Dinucleotide composition-dependent DNA structural features strongly influence aspects of protein–DNA interactions. Certain base steps are associated with an intrinsic curvature, which can lead to bending and supercoiling, and might influence nucleosome placements[20,23]. The orientation of a DNA molecule relative to a protein surface is determined mainly by the directional bending preferences of the DNA, rather than by any sequence-specific protein–DNA contacts[20]. Many DNA-repair enzymes recognize shapes and lesions in DNA structures, rather than specific sequences; the topological state of the DNA can affect the rate of damage recognition and repair[25,26]. DNA structures may be crucial in modulating the processes of replication and repair. There appear to be biases in the replication efficiency and fidelity, depending on neighboring base context[25,26], which is putatively related to stacking capacities. The interaction of nucleosome position with DNA-binding proteins, and ribosomal binding of mRNA, are strongly affected by dinucleotide arrangements[20].

Perspectives

Genomic sequences display heterogeneity on many scales. Many authors have emphasized variation in G+C content, e.g. isochore compartments. The dinucleotide content is of equal or greater importance. DNA structural configurations appear to be determined principally by base-step arrangements. In fact, observation of the distribution of dinucleotides separated by 0, 1, 2 or more nucleotides (Fig. 1) has shown that, although dinucleotide relative abundance values for 0 gap are in many cases highly biased and putatively important

Box 1. Genomes studied

Prokaryotes		
Gram-negative	α-Proteobacteria	*Agrobacterium tumefaciens, Paracoccus denitrificans, Rhodobacter capsulatus, Rhodobacter sphaeroides, Rhizobium meliloti*
	β-Proteobacteria	*Neisseria gonorrhoeae*
	γ-Proteobacteria	*Azotobacter vinelandii, Pseudomonas aeruginosa, Haemophilus influenzae, Klebsiella pneumoniae, Escherichia coli, Salmonella typhimurium*
	δ-Proteobacteria	*Myxococcus xanthus*
Gram-positive		*Bacillus subtilis, Bacillus stearothermophilus, Lactococcus lactis, Mycobacterium leprae, Mycobacterium tuberculosis, Staphylococcus aureus, Streptomyces griseus, Streptomyces lividans*
Mycoplasma		*Mycoplasma capricolum, Mycoplasma pneumoniae*
Cyanobacteria		*Anabaena* spp., *Synechococcus* spp.
Spirochetes		*Borrelia burgdorferi, Treponema* spp., *Leptospira* spp.
Unassigned		*Thermus* spp., *Thermotoga* spp.
Archaebacteria		*Halobacterium halobium, Methanococcus* spp., *Methanobacterium thermoautotrophican, Sulfolobus* spp., *Pyrococcus* spp.
Phage	(ds) Temperate	λ, Mu, P1, P4, P22
	(ds) Lytic	T2, T4, T3, T7, φ29
	Parasitic-filamentous	f1, I22, IKe, PF3
	ssDNA	φX174, G4
	ssRNA	MS2, GA
Organelles		
Mitochondria (complete genomes)	Metozoa	12 vertebrates, 7 invertebrates
	Fungi	*Saccharomyces cerevisiae, Schizosaccharomyces pombe, Podospora anserina*
	Protist	*Trypanosoma brucei, Paramecium aurelia*
	Plant	Liverwort
Chloroplasts (complete genomes)		Rice, tobacco, *Euglena gracilis*, liverwort, *Epifagus virginia*
Eukaryotes		
Vertebrates		Human, bovine, sheep, pig, rabbit, dog, mouse, rat, chicken, *Xenopus laevis*, trout
Invertebrates		*Drosophila melanogastor, Drosophila pseudoobscura, Manduca sexta, Bombyx mori, Caenorhabditis elegans, Strongylocentrotus purpuratus*
Fungi		*Saccharomyces cerevisiae, Schizosaccharomyces pombe, Neurospora crassa, Aspergillus nidulans*
Protists		*Giardia lambdia, Trypanosoma brucei, Leishmania* spp., *Entamoeba histolytica, Plasmodium falciparum, Dictyostelium discoideum, Tetrahymena* spp., *Acanthamoeba* spp., *Trichomonas* spp., *Toxoplasma* spp.
Plants	Monocots	Maize, rice, barley (*Hordeum vulgare*), wheat (*Triticum aestivum*)
	Dicots	*Arabidopsis thaliana*, soybean, tobacco (*Nicotiana tabacum*), *Antirrhinum majus, Brassica napus*, cotton (*Gossypium hirsutum*), alfalfa (*Medicago sativa*)
Vertebrate small viruses (≤30 kb; all complete genomes)		Papova- and papillomaviruses (10 complete genomes), hepadna (6), parvo (4), lenti (7), other retro (11), toga (10), picorna (10), flavi (9), calici (4), corona (1), orthomyxo (3), rhabdo (2) viruses
Vertebrate large viruses (>30 kb; all complete genomes)		Adeno (2), herpes (8), vaccinia (1) viruses
Plant viruses		33 RNA, 3 DNA

investigation of the energy minima for the geometry of contiguous base pairs in terms of slide, roll, tilt and helical-twist parameters suggests that the ten symmetric dinucleotides largely account for the DNA structures observed with X-ray diffraction of special synthesized oligonucleotides[22,24,27]. Dinucleotide properties, including stacking energies, charge interactions and conformational tendencies, are paramount in determining local DNA structure[20,22–24]. For example, whenever a DNA molecule is bent, the resultant DNA configuration reflects local distortions in the double helix at the level of individual dinucleotide steps[20,23].

In view of the limited nature of dinucleotide relative abundance values within genomes, we propose the set of all symmetric, dinucleotide relative abundance values as a signature for discriminating genomic DNA. Figure 2 shows the range of dinucleotide relative abundance values based on disjoint 100 kb, 50 kb and 20 kb samples obtained from partitioning a 1.25 Mb *E. coli* contig (extending from 67° to 4°) and from a

1.06 Mb contig from *Caenorhabditis elegans*. This figure demonstrates the stability of the signature within a genome and its capacity to distinguish between sequences from different organisms (e.g. GC is persistently over-represented in *E. coli* but normal in *C. elegans*, and *C. elegans* is more potently TA suppressed compared with *E. coli*).

In discriminating between classes of organisms, we emphasize the presence or absence of dinucleotide relative abundance extremes in the signature (Table 2). This perspective was applied to propose a different bacterial ancestor for animal mt DNA from the classical ones[13]. Comparing animal mt genomes against a broad spectrum of bacterial sequences, we found the animal mt signature to be in complete accord with that of *Sulfolobus* spp. and substantially in accord with *M. capricolum*. In contrast, the animal mitochondrial signature is very different from those of α- and γ-proteobacteria. Another intriguing observation is that the genomic signature of vertebrates (but not of invertebrates) is in substantial agreement with the genomic signature of thermophilic Archaea, but not similar to the signature of sequences from halophilic Archaea or to any eubacterial species[5,7,9]. This similarity may relate to the effects of the higher temperatures experienced by the genomes of thermophilic bacteria and (warm-blooded) vertebrates.

The genome-wide consistency of dinucleotide relative abundance values suggests involvement of genome-wide processes, such as replication, recombination and repair. Environmental influences on the DNA sequence and concomitantly on the genomic signature include UV-radiation damage, osmolarity gradients, temperature extremes, acidity or alkalinity, ecology (e.g. energy sources and systems) and direct, or indirect, transfer of genomic DNA between organisms.

What are the possible mechanisms underlying the signature determinations? High CC/GG levels [often higher than TG/CA values (Table 2)] frequently correlate with CG suppression. These data argue that non-methylation mechanisms frequently underlie the suppression of CG. These mechanisms are putatively related to DNA structure and are probably influenced by interspecific differences in replication and repair systems[25,26]. Is it possible that CG dinucleotides are important components of regulatory or structural sequences whose frequency should be kept distinctly low for optimum functioning? From another perspective, wherever CG is a 'hot spot' of mutation with potentially deleterious functional, or structural, consequences for DNA and proteins, a reduction in CG occurrence would be selectively favorable. In dicot plants, we observed CG suppression, CC/GG of normal relative abundance, TG/CA marginally high and CNG not suppressed. The DNA of monocot grasses is substantially methylated (≥25% of C nucleotides), to a similar extent to vertebrate DNA, but are not CG suppressed. Obviously, mechanisms underlying CG representations are varied and not well understood. In contrast, the nearly universal under-representation of TA is probably a consequence of the unusually low stacking energy and the unique conformational preferences of this dinucleotide.

Acknowledgements

We thank Drs B.E. Blaisdell, V. Brendel, A.M. Campbell and J. Eisen for helpful comments on the manuscript. S.K. and C.B. were supported, in part, by NIH grants 5R01GM10452-30, 5R01HG0035-07 and NSF grant DMS 9403553.

References

1 Josse, J., Kaiser, A.D. and Kornberg, A. (1961) *J. Biol. Chem.* 236, 864–875
2 Swartz, M.N., Trautner, T.A. and Kornberg, A. (1962) *J. Biol. Chem.* 237, 1961–1967
3 Russel, G.J., Walker, P.M.B., Elton, R.A. and Subak-Sharpe, J.H. (1976) *J. Mol. Biol.* 108, 1–28
4 Russel, G.J. and Subak-Sharpe, J.H. (1977) *Nature* 266, 533–535
5 Karlin, S., Ladunga, I. and Blaisdell, B.E. (1994) *Proc. Natl Acad. Sci. USA* 91, 12837–12841
6 Karlin, S., Mocarski, E.S. and Schachtel, G.A. (1994) *J. Virol.* 68, 1886–1902
7 Karlin, S. and Ladunga, I. (1994) *Proc. Natl Acad. Sci. USA* 91, 12832–12836
8 Burge, C., Campbell, A.M. and Karlin, S. (1992) *Proc. Natl Acad. Sci. USA* 89, 1358–1362
9 Karlin, S. and Cardon, L.R. (1994) *Annu. Rev. Microbiol.* 44, 619–654
10 Doerfler, W. (1983) *Annu. Rev. Biochem.* 52, 93–124
11 Bestor, T.H. and Coxon, A. (1993) *Curr. Biol.* 6, 384–386
12 Cardon, L.R., Burge, C., Clayton, D.A. and Karlin, S. (1994) *Proc. Natl Acad. Sci. USA* 91, 3799–3803
13 Karlin, S. and Campbell, A.M. (1994) *Proc. Natl Acad. Sci. USA* 91, 12842–12836.
14 Karlin, S., Doerfler, W. and Cardon, L.R. (1994) *J. Virol.* 68, 2889–2897
15 Shpaer, E.G. and Mullins, J.I. (1990) *Nucleic Acids Res.* 18, 5793–5797
16 Drake, J.W. *et al.* (1969) *Nature* 221, 1128–1132
17 Breslauer, K.J., Frank, R., Blöcker, H. and Marky, L.A. (1986) *Proc. Natl Acad. Sci. USA* 83, 3746–3750
18 Delcourt, S.G. and Blake, R.D. (1991) *J. Biol. Chem.* 266, 15160–15169
19 Finnegan, E.J., Brettell, R.I.S. and Dennis, E.S. (1993) in *DNA Methylation, Molecular Biology and Biological Significance* (Jost, J.P. and Saluz, H.P., eds), pp. 218–261, Birkhäuser Verlag
20 Travers, A. (1993) *DNA–Protein Interactions*, Chapman & Hall
21 Beutler, E. *et al.* (1989) *Proc. Natl Acad. Sci. USA* 86, 192–196
22 Hunter, C.A. (1993) *J. Mol. Biol.* 230, 1025–1054
23 Calladine, C.R. and Drew, H.R. (1992) *Understanding DNA*, Academic Press
24 Quintana, J.R., Yangi, K. and Dickerson, R.E. (1992) *J. Mol. Biol.* 225, 379–395
25 Echols, H. and Goodman, M.F. (1991) *Annu. Rev. Biochem.* 60, 477–511
26 Kunkel, T.A. (1992) *Bioessays* 14, 303–308
27 Dickerson, R.E. (1992) *Methods Enzymol.* 211, 67–111

S. KARLIN AND C. BURGE ARE IN THE DEPARTMENT OF MATHEMATICS, STANFORD UNIVERSITY, STANFORD, CA 94305-2125, USA.