Annu. Rev. Genet. 1998. 32:185–225

COMPARATIVE DNA ANALYSIS ACROSS DIVERSE GENOMES

Samuel Karlin[1], *Allan M. Campbell*[2], *and Jan Mrázek*[1]
[1]Department of Mathematics and [2]Department of Biological Sciences, Stanford University, Stanford, California 94305-2125

KEY WORDS: genome signature, compositional biases, frequent oligonucleotides (motifs), strand compositional asymmetry, codon biases

ABSTRACT

We review concepts and methods for comparative analysis of complete genomes including assessments of genomic compositional contrasts based on dinucleotide and tetranucleotide relative abundance values, identifications of rare and frequent oligonucleotides, evaluations and interpretations of codon biases in several large prokaryotic genomes, and characterizations of compositional asymmetry between the two DNA strands in certain bacterial genomes. The discussion also covers means for identifying alien (e.g. laterally transferred) genes and detecting potential specialization islands in bacterial genomes.

CONTENTS

0066-4197/98/1215-0185$08.00

INTRODUCTION

Molecular sequence data are accumulating at an unprecedented pace. Dozens of complete genomes, tens of thousands of proteins, and several hundred nonredundant protein structures are now available. The coming phase of molecular biology will see increasing efforts to categorize and analyze these data using empirical and interactive statistical and computational methods with the goal of understanding on a molecular level the nature of information: its mode of expression and its biological meaning, its transfer in biological systems, and its evolution.

Genomic global and local compositional heterogeneity is widely recognized. The many facets of DNA heterogeneity include isochore compartments in vertebrate species (5) and the G+C- and A+T-rich halves of the bacteriophage lambda genome (40); transposable elements (such as Ty in yeast, IS in *Escherichia coli* and Alu in human) (4); centromeric satellite tandem repeats (such as the 171-bp human alpha satellite DNA) (92); characteristic telomeric sequences (such as the hexanucleotide AGGGTT tandem repeats in humans) (9); repetitive extragenic palindromes (REPs) of *E. coli* and *Salmonella typhimurium* (11, 32, 61); repeat induced point mutation (RIP) in *Neurospora* and other fungi (83); recombinational hot spots [such as chi elements in *E. coli* (61)]; universal underrepresentation of the dinucleotide TpA (20); underrepresentation of the dinucleotide CpG in vertebrates and many thermophiles (41, 56); HTF islands (DNA sequences that generally occur upstream of vertebrate genes and are abundant with nonmethylated CpG) (8); the underrepresentation of the tetranucleotide CTAG in proteobacterial genomes (45, 56); GNN periodicity in coding sequences (27); and methyltransferase modifications (73). Thus, genome organization is complex and variable. In particular, eukaryotic sequences are often endowed with tandem repeats accruing from polymerase slippage or unequal crossing-over and with distant direct and inverted repeats promoted in part by transposition, translocation, recombination, amplification, and excision. Many genomic sequences exhibit polymorphisms, strain variation, DNA inversions, and rearrangements reflecting a state of flux.

Prokaryotic genomes especially are in a state of flux influenced by natural genetic transformation (competence). Under appropriate conditions, almost all cells of *Haemophilus influenzae* and *Neisseria gonorrhoeae* are competent. Generally, although exogenous DNA incorporation is widespread in bacterial cells, nonspecific integration into the chromosome seems to be rare (69). Biological phenomena are generally highly variable at the molecular level, a circumstance enabling evolutionary developments. [See the discussion between the protagonists (58) and antagonists (31) of the neutral theory of molecular evolution for explanations of the extant variability.]

Since 1995 more than two dozen complete prokaryotic and eukaryotic genomes have been reported and many more genomes and chromosomal sets are forthcoming. These genomes provide opportunities and pose challenges for characterizing genomic inhomogeneities, for detecting significant sequence patterns, and for evolutionary comparisons unbiased by selective sequencing. The first step of genome analysis commonly aims to identify the gene repertoire emphasizing similiarities, differences, and uniqueness among genes (e.g. 60, 89). These authors have introduced methods to determine metabolic pathways exploiting comparative functional genomics. A caveat: Sequence (or gene) similarity does not per se imply functional/structural concordance and sequence differences do not per se preclude similar function (for examples, see 88).

Methods for analyzing genomes emphasizing sequence features other than gene comparisons rely on the following assessments of genomic organization and sequence heterogeneity: (*a*) compositional biases of short oligonucleotides; (*b*) dinucleotide relative abundances (the genome signature); (*c*) codon and residue biases; (*d*) rare and frequent words (oligonucleotides, peptides, codons); (*e*) clustering, overdispersion, or excessive evenness in the distribution of various markers, e.g. particular oligonucleotides, restriction sites, nucleosome placements, methylation targets, origins of replication, repair recognition sites, a myriad of control sequences; and (*f*) repeat structures in the genome.

This review emphasizes four (interrelated) areas:

1. Genomic signatures and their evolutionary implications. In particular, we apply the dinucleotide relative abundance profile for genome comparisons and phylogenetic reconstructions that do not require alignment. DNA structure and evolution is fundamental for understanding biases in dinucleotide relative abundance profiles (the genomic signature).

2. Statistical methods for genome analysis. In this context the use of r-scan statistics affords means to assess anomalies in the distribution of specific

markers along sequences and characterizations of genomic heterogeneity within and between species (e.g. rare and frequent words, motifs or compositional biases).

3. Genomic codon usage patterns. Identification of constraints on codon and amino acid usages, codon bias, and genomic signature fluctuations help in detecting potential pathogenicity islands and in identifying laterally transferred genes.

4. Strand compositional asymmetry. Data are presented and interpretations are proffered in terms of replication asymmetries, mutational biases, transcription coupled repair mechanisms, and concomitants of multiple origins of replication.

GENOME SIGNATURE

Dinucleotide relative abundance values (dinucleotide bias) are assessed 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 corresponding frequencies of the sequence concatenated with its inverted complementary sequence (44, 56). Dinucleotide relative abundance profiles $\{\rho^*_{XY}\}$ differences from 1 effectively assess contrasts between the observed dinucleotide frequencies and those expected from random associations of the component mononucleotide frequencies. From data simulations and statistical theory, estimates of $\rho^*_{XY} \leq 0.78$ or $\rho^*_{XY} \geq 1.23$ convey significant underrepresentation or overrepresentation, respectively, for sufficiently long (say $\geq$50 kb) random sequences, with the probability at most 0.001 for observing such an extreme base composition. For a random sequence, ρ^*_{XY} values, for all XY, approach 1 (deviation from 1 is about $1/\sqrt{n}$ for sequences of length n). Therefore, for $n \sim 100,000$, $|\rho^*_{XY} - 1|$ is of the order 0.003.

The dinucleotide relative abundance values (Table 1) evaluated for ($\geq$50 kb) multiple DNA contigs from the same organism are generally much more similar to each other than they are for sequence contigs from different organisms (see below), and closely related organisms generally have more similar dinucleotide relative abundance values than do distantly related organisms (44, 49, 56). Dinucleotide relative abundance values are equivalent to the robust "general designs" derived from biochemical nearest-neighbor frequency analysis (41, 80, 81). These highly stable DNA doublets are essentially constant in most organisms for bulk DNA including protein coding DNA and for DNA fractions of differing sequence complexity (81), suggesting that there may be genome-wide factors such as functions of the replication and repair

machinery, context-dependent mutations rates, DNA modifications, and base-step conformational tendencies that impose limits on the compositional and structural patterns of a genome sequence. Thus, the set of dinucleotide relative abundance values constitutes a genomic signature (44, 56) that may reflect the influence of such factors.

Dinucleotide relative abundances capture most of the departure from randomness in genome sequences. Comparisons were made in terms of di-, tri-, and tetranucleotide relative abundance differences. The di- and the corresponding di- + tri- + tetra-relative abundances between sequences correlate highly (47, 49), suggesting that DNA conformational arrangements are principally determined by base-step configurations (16, 24). Analysis of the distribution of dinucleotide relative abundances separated by $k = 1, 2, \ldots,$ K other nucleotides has shown that although values for no separation are often highly biased, those for separation by one or more nucleotides are more nearly random (44). More specifically, $\rho^*(XN_kY), k \geq 1$ are almost always in the random range and uninformative. Parenthetically, prokaryotic genomes tend to be homogeneous in their G+C content but this property is not diagnostic in discriminating among prokaryotes.

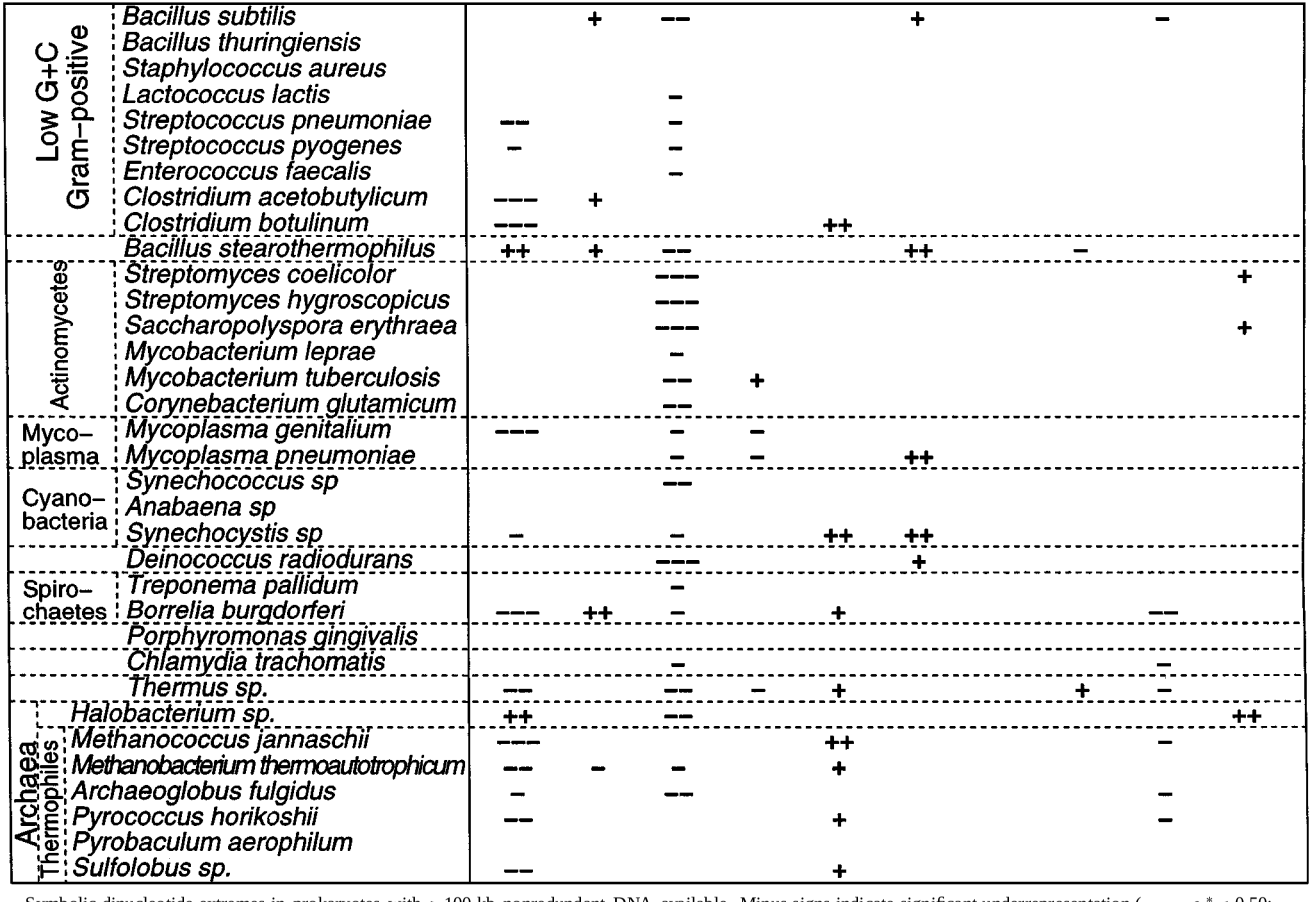
Comparisons Among Genome Signature Values

CG is underrepresentated (significantly low relative abundances) in vertebrate sequences, many protist genomes (*Plasmodium falciparum, Dictiostelium discoideum, Entamoeba histolytica*; but not *Trypanosoma brucei*), dicots (44), animal mitochondrial genomes (22), small viral genomes (48), several thermophilic bacteria (56), and several prokaryotic species, e.g. *Borrelia burgdorferi, Clostridium acetobutylicum,* and *Mycoplasma genitalium,* and overrepresented in *Halobacterial* sp., *Bacillus stearothermophilus,* and *Neisseria gonorrhoeae* (53). The dinucleotide TA is broadly underrepresented in the bulk of prokaryotic and eukaryotic sequences (54, 56). In contrast, TA representations are normal in *C. acetobutylicum* and in the archaeal genomes of *Pyrococcus horikoshii, Pyrobaculum aerophilum, Methanococcus jannaschii,* and also in *Sulfolobus* sp.

The two spirochaetes *T. pallidum* vs *B. burgdorferi* sharply contrast in ρ^*_{XY} extremes for CG, GC, CC/GG, and AC/GT. The CG representations of *M. genitalium* and *Mycoplasma pneumoniae* clearly deviate but have close relative abundance extremes for TA, AT, and TT/AA dinucleotides. *M. jannaschii, M. thermoautotrophicum,* and *Archaeoglobus fulgidus* differ much in their ρ^*_{XY} profiles. Notable contrasts: *C. acetobutylicum* is significantly underrepresented in CG but significantly overrepresented in GC. *Mycobacterium tuberculosis* is significantly low in TA and significantly high in AT. The archaeal genome *P. aerophilum* is normal in all dinucleotide relative abundances.

Table 1 Symbolic dinucleotide extremes in prokaryotes

			CG	GC	TA	AT	CC GG	TT AA	TG CA	AG CT	AC GT	GA TC
Gram–negative proteobacteria	γ	*Escherichia coli*		+	–							
		Salmonella typhimurium		+								
		Klebsiella pneumoniae		++		+					–	
		Yersinia enterocolitica		+								
		Vibrio cholerae		+								
		Haemophilus influenzae		++	–			+				
		Coxiella burnetii		+	–			+			–	
		Buchnera aphidicola										
		Pseudomonas putida			––							
		Pseudomonas aeruginosa			––							
		Azotobacter vinelandii			––							
	β	*Bordetella pertussis*		+	––	++						
		Alcaligenes eutrophus		+	––	++						
		Neisseria gonorrhoeae	++	+	––			++		––		
		Neisseria meningitidis	++	+	––			++		–		
	α	*Agrobacterium tumefaciens*			––							
		Rhizobium leguminosarum	+		––	++						
		Rhizobium meliloti	+		––	++						+
		Bradyrhizobium japonicum	+		–––	++	–					
		Rhodobacter capsulatus			–––	+++		+			–	
		Rhodobacter sphaeroides			–––	++						+
		Paracoccus denitrificans			–––	+++					–	
		Rickettsia prowazekii		++								
	δ	*Myxococcus xanthus*			–––							
	ε	*Helicobacter pylori*		+++	–			++			––	



Symbolic dinucleotide extremes in prokaryotes with >100 kb nonredundant DNA available. Minus signs indicate significant underrepresentation ($---\ \rho^* < 0.50$; $--\ 0.50 < \rho^* < 0.70$; $-\ 0.70 < \rho^* < 0.78$), plus signs indicate significant overrepresentation ($+++\ \rho^* > 1.50$; $++\ 1.30 < \rho^* < 1.50$; $+\ 1.23 < \rho^* < 1.30$). No symbol indicates ρ^* in the normal range. The common taxonomic classification is indicated.

TT/AA is overrepresented in several proteobacteria, *Mycoplasmas, Synechocystis*, and *Deinococcus radiodurans* among eubacteria. There are no underrepresentations of TT/AA. High representations of CC/GG include *Synechosystis, B. burgdorferi, M. jannaschii, M. thermoautotrophicum*, and *P. horikoshii*. The symmetric dinucleotide relative abundances TG/CA and GA/TC are pervasively in the normal range (the same for AG/CT except for the *Neisseria* genomes). The dinucleotide AT predominantly shows normal representations except for *Mycoplasma* (low) and *M. tuberculosis* (high).

Dinucleotide Compositional Extremes in Prokaryotic Genomes

Table 1 summarizes the dinucleotide relative abundance extremes for an updated list of sequence collections. The limited range of the ρ^*_{XY} values over multiple 50-kb contigs [consult (53, 54, 56)] confirms the substantial invariance of the dinucleotide relative abundance profile. (The results are even more stable for larger contig size, e.g. 100 kb.) There are clear trends, as follows.

1. The dinucleotide TA is broadly underrepresented or low normal in prokaryotic sequences at the level $0.50 \leq \rho^*_{TA} \leq 0.82$ (exceptions include the two archaea *P. aerophilum* ($\rho^*_{TA} \sim 1.07$) and *Sulfolobus* sp. ($\rho^*_{TA} \sim 1.01$)] (47, 56). TA underrepresentation is also pervasive in eukaryotic chromosomes but not in eukaryotic viral genomes or in organellar genomes (22, 46).

2. GC is predominantly overrepresented in γ-proteobacterial sequences, in many β-proteobacterium examples, and in several low-G+C Gram-positive bacterial genomes (e.g. *B. subtilis* and *C. acetobutylicum*).

3. CG is underrepresented in *M. genitalium* (but not in *M. pneumoniae*) and in the low-G+C Gram-positive sequences of *Streptococcus* and *Clostridium* and in many thermophiles, including *M. jannaschii, Sulfolobus* sp., *M. thermoautotrophicum*, and *Thermus* sp., but not in *P. aerophilum* or *P. horikoshii*. At the other extreme, CG is overrepresented in *Bacillus stearothermophilus*, in halophiles, and also in several β- and α-proteobacterial genomes (e.g. *Rhizobium* sp. and *Neisseria* sp.).

4. AT is overrepresented in most α-proteobacterial sequences.

5. Only a few bacterial genomic sequences are devoid of any dinucleotide extremes. All dinucleotide relative abundances are in the random range for *S. aureus*, *Anabaena*, and *P. aerophilum* (Table 1).

Dinucleotide Compositional Extremes in Eukaryotic Genomes

The following trends were observed.

1. TA is broadly underrepresented in eukaryotic chromosomes generally in the range $\rho^*_{TA} \sim 0.61$–0.81. TA occurrences are in the random (normal) range in animal mitochondrial (Mt) sets and chloroplast genomes. Possible reasons for TA underrepresentation may be its low thermodynamic stacking energy, which is the lowest among all dinucleotides (e.g. 16, 24), the high degree of degradation of UA dinucleotides by ribonucleases in mRNA tracts (6), and the presence of TA as part of many regulatory signals (e.g. TATA box, transcription terminators). From this perspective, TA suppression may help to avoid inappropriate binding of regulatory factors.

2. CG shows drastic suppression in vertebrates. Overall, ρ^*_{CG} values in vertebrates range from 0.23 to 0.37. CG is strongly suppressed in the sea urchin *Strongylocentrotus purpuratus* (0.59), in some yeasts (*Kluyveromyces lactis* and *Candida albicans*), and in dicot plants, but is only marginally low to low normal in monocot plants (44). CG is suppressed in animal mitochondria (ρ^* values mostly in the range 0.50–0.65), whereas it is in the normal range in higher plant chloroplast genomes (46). 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 (25, 90). 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 $\rightarrow$ CC/GG mutation bias. We have proposed that CG deficiencies may in some circumstances be selected because of structural constraints related to high dinucleotide stacking energy, supercoiling, and chromatin packing (44).

3. The dinucleotides CC/GG, TG/CA, and AG/CT, all a single-base mutation from CG, are (except for dicot plants) overrepresented only in genomes with strong CG suppression. These dinucleotide relative abundances separate rodents, posssessing 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 (Table 2) (54).

Table 2 Symbolic dinucleotide extremes in eukaryotes

			CG	GC	TA	CC GG	TT AA	TG CA	AG CT
Deuterostomes	Vertebrates	human	−−−		−	+			
		cow	−−−		−				
		pig	−−−		−−	+			
		rabbit	−−−		−	+			
		mouse	−−−		−			+	+
		rat	−−−		−			+	+
		hamster	−−−		−−			+	++
		chicken	−−−		−−			++	+
		Xenopus laevis	−−−						
	S. purpuratus		−−		−−				
Proto–stomes	*Drosophila melanogaster*			+	−		+		
	Drosophila virilis			++					
	Bombyx mori								
	Caenorhabditis elegans				−−		+		
Fungi	*S. cerevisiae*				−				
	Kluyveromyces lactis		−						
	Candida albicans		−−		−−			+	
	S. pombe				−				
	Neurospora crassa				−−				
	Emericella nidulans				−				
	Aspergillus niger				−				
	Ustilago maydis				−−	−			
Plants	*Arabidopsis thaliana*		−		−				
	tobacco		−−		−				
	potato		−−						
	tomato		−−						
	maize		−						
	barley								
	rice		−		−				
Protists	*Plasmodium falciparum*		−			++			
	Trypanosoma brucei				−				
	Dictiostelium discoideum		−		−−				

Symbolic dinucleotide extremes in eukaryotes with >100 kb nonredundant DNA available. The listed eukaryotes exhibit no significant extremes for dinucleotides AT, AC/GT and GA/TC. See also legend to Table 1.

4. Other dinucleotide biases in eukaryotes include overrepresentation of GC in *Drosphilia* species but apparently not in other higher eukaryotes. GC is significantly abundant in most γ-proteobacteria (56).

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

CODON SIGNATURE

For a given collection of genes, let $f_X(1)$, $f_Y(2)$, $f_Z(3)$ denote frequencies of the indicated nucleotide at codon sites 1, 2, and 3, respectively, and let f_{XYZ} indicate codon frequency. The embedded dinucleotide frequencies are denoted $f_{XY}(1,2)$, $f_{YZ}(2,3)$, and $f_{XZ}(1,3)$. Dinucleotide contrasts are assessed through the odds ratio $\rho_{XY} = f_{XY}/f_X f_Y$. In the context of codons, we define

$$\rho_{XY}(1,2) = \frac{f_{XY}(1,2)}{f_X(1)\, f_Y(2)},$$

$$\rho_{YZ}(2,3) = \frac{f_{YZ}(2,3)}{f_Y(2)\, f_Z(3)},$$

$$\rho_{XZ}(1,3) = \frac{f_{XZ}(1,3)}{f_X(1)\, f_Z(3)}.$$

We refer to the profiles $\{\rho_{XY}(1,2)\}$, $\{\rho_{XZ}(1,3)\}$, $\{\rho_{YZ}(2,3)\}$, and also $\{\rho_{ZW}(3,4)\}$, where 4 (=1) is the first position of the next codon, as the codon signature to be distinguished from the global genome signature (52).

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. CG and TA suppression in human (and vertebrate) sequences is a strong component of the dinucleotide biases in all coding and noncoding sequences of human. CG suppression is stronger in noncoding sequences, whereas TA suppression is stronger in genes, perhaps because of high susceptibility of RNase activity in transcripts containing UA (6). CG is less suppressed at sites {1, 2}, probably reflecting requirements of Arg usage (52).

In human sequences, even though G is the most frequent nucleotide (32–33%), at codon site $1 = 4$ and C is the most frequent nucleotide at codon site 3 (29.3%), the dinucleotide CG frequency is significantly deficient. Moreover, the extent of CG suppression is less extreme at codon junctions ($\rho_{CG}(3,4) \approx 0.44$) compared to codon positions {2, 3} ($\rho_{CG}(2,3) \approx 0.36$) within a codon. One way to explain this inequality recognizes the methylation/deamination/

mutation pathway coupled to the hypothesis that DNA repair in the transcribed strand is more proficient than in the nontranscribed strand (36). Specifically, comparing CG at {2, 3} with CG at {3, 4}, we assume that the methylation/deamination/mutation scenario creates mutation at nucleotide C much more than at nucleotide G.

It is of interest to compare the codon signature with the genome signature. The genome and codon signatures of human are qualitatively concordant (52). This result is consistent with our thesis that codon choice in human (and mammalian) genes is largely a consequence of two factors: (*a*) constraints on amino-acid usages essential for protein structure/function; and (*b*) maintaining DNA structures dependent on base-step conformational tendencies consistent with the organism's genome signature determined by genome-wide processes of DNA modification, replication, and repair (52).

MEASURES OF DIFFERENCES WITHIN AND BETWEEN GENOMES

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

$$\delta^*(f, g) = 1/16 \sum_{XY} |\rho^*_{XY}(f) - \rho^*_{XY}(g)|,$$

where the sum extends over all dinucleotides (abbreviated δ^*-differences). Table 3 compares δ^* (f, g) values within and between large genomic sequence sets. The average δ^*-differences are based on multiple 50-kb contigs. To avoid the possibility of a few extreme dinucleotide relative abundances exerting a large influence on the δ^*-value, we have introduced a method of partial orderings comparing the complete genome signature vector of the two sequences. The partial orderings are consistent with accepted evolutionary relationships and reinforce our conclusions from the distance analysis. For rationale, precision, and examples, see (49, 51, 56).

Figure 1 displays a set of histograms generated by all pairwise δ^*-differences among nonoverlapping 50 kb contigs of selected species. For convenience, we describe levels of δ^*-differences for some reference examples (all values mutliplied by 1000):

Close ($\delta^* \leq 50$; pervasively within species, human vs cow, *Lactococcus lactis* vs *Streptococcus pyogenes*).

Moderately similar ($55 \leq \delta^* \leq 85$; human vs chicken, *Escherichia coli* vs *Haemophilus influenzae*, *Synechococcus* vs *Anabaena*).

Weakly similar ($90 \leq \delta^* \leq 120$; human vs sea urchin, *M. genitalium* vs *M. pneumoniae*).

Distantly similar ($125 \leq \delta^* \leq 145$; human vs *Sulfolobus*, *E. coli* vs *R. prowazekii*, *M. jannaschii* vs *M. thermoautotrophicum*).

Distant ($150 \leq \delta^* \leq 180$; human vs *Drosophilia*, *E. coli* vs *Helicobacter pylori*).

Very distant ($\delta^* \geq 190$; human vs *E. coli*, *E. coli* vs *Sulfolobus*, *M. jannaschii* vs *Halobacterium*).

Within-species δ^*-differences (diagonal elements of Table 3) range from 20–43 (all δ^*-differences are multiplied by 1000), whereas the average between-species δ^*-differences range from 34–309. Thus, within-species δ^*-differences are persistently of lower values compared to between-species.

Prokaryotic Taxonomy

There are many uncertainties and active debates regarding the taxonomy of prokaryotes [for a recent review see (17)]. It is of interest to see how genomic signature information correlates with other measures of sequence similarity.

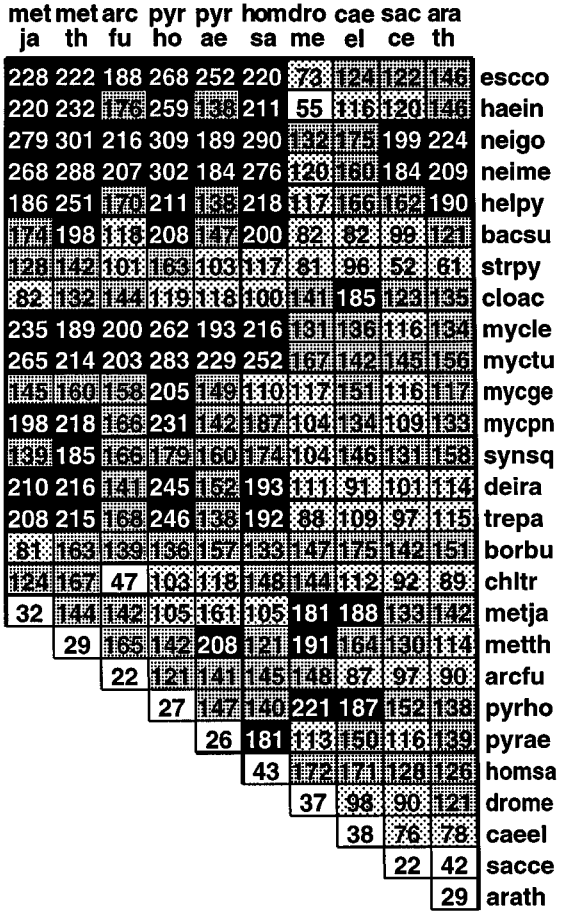
Table 3 Average δ^*-differences based on 50 kb sequence samples (values multiplied by 1000)

esc co	hae in	nei go	nei me	hel py	bac su	str py	clo ac	myc le	myc tu	myc ge	mycsyn pn	dei sq	tre ra	tre pa	bor bu	chl tr	
26	60	115	95	173	87	114	187	85	107	158	150	153	94	66	197	175	escco
	28	115	103	124	90	109	176	129	158	143	126	128	112	89	166	169	haein
		23	34	165	139	188	237	190	188	230	190	169	172	164	244	217	neigo
			31	161	123	173	225	169	169	219	181	162	152	143	232	208	neime
				29	140	148	165	238	271	168	130	131	178	170	111	160	helpy
					38	103	167	122	136	166	152	145	89	94	149	119	bacsu
						27	102	139	177	86	109	138	95	90	111	90	strpy
							24	207	245	119	169	151	191	173	87	126	cloac
								18	53	187	166	195	102	79	242	185	mycle
									26	228	202	226	117	114	277	206	myctu
										43	99	159	140	131	122	161	mycge
											38	120	134	122	171	168	mycpn
												24	187	173	138	175	synsq
													20	57	185	150	deira
														26	188	152	trepa
															25	133	borbu
																24	chltr

List of species: escco (*Escherichia coli*, 4.64Mb aggregate sequence sample), haein (*Haemophilus influenzae*, 1.83Mb), neigo (*Neisseria gonorrhoeae*, 878kb), neime (*Neisseria meningitidis*, 2.21Mb), helpy (*Helicobacter pylori*, 1.67Mb), bacsu (*Bacillus subtilis*, 4.21Mb), strpy (*Streptococcus pyogenes*, 985kb), cloac (*Clostridium acetobutylicum*, 4.03Mb), mycle (*Mycobacterium leprae*,1.68Mb), myctu (*Mycobacterium tuberculosis*, 4.41Mb), mycge (*Mycoplasma genitalium*, 580kb), mycpn (*Mycoplasma pneumoniae*, 816kb), synsq (*Synechocystis* sp., 3.57Mb), deira (*Deinococcus radiodurans*, 3.06Mb), trepa (*Treponema pallidum*, 1.14Mb), borbu (*Borrelia burgdorferi*, 911kb), chltr (*Chlamydia trachomatis*, 1.04Mb), metja (*Methanococcus jannaschii*, 1.66Mb), metth (*Methanobacterium thermoautotrophicum*, 1.75Mb), arcfu (*Archaeoglobus fulgidus*2.18Mb), pyrho (*Pyrococcus horikoshii*, 1.09Mb), pyrae (*Pyrobaculum aerophilum*, 2.17Mb), homsa (human, 5.84Mb), drome (*Drosophila melanogaster*, 4.30Mb), caeel (*Caenorhabditis elegans*, 7.10Mb), sacce (yeast, 12.0Mb), arath (*Arabidopsis thaliana*, 1.99Mb).

(*Continued*)

Table 3 (*Continued*)



Diagonal entries show average within-species δ^*-differences (average over all pairwise comparisons between disjoint 50 kb samples from the same genome). Non-diagonal entries show average between-species δ^*-differences. The values of $\delta^* \leq 60$ are printed on a white background, $60 < \delta^* \leq 120$ on a light dotted background, $120 < \delta^* \leq 180$ on a dense dotted background and values of $\delta^* > 180$ are printed white on black background.

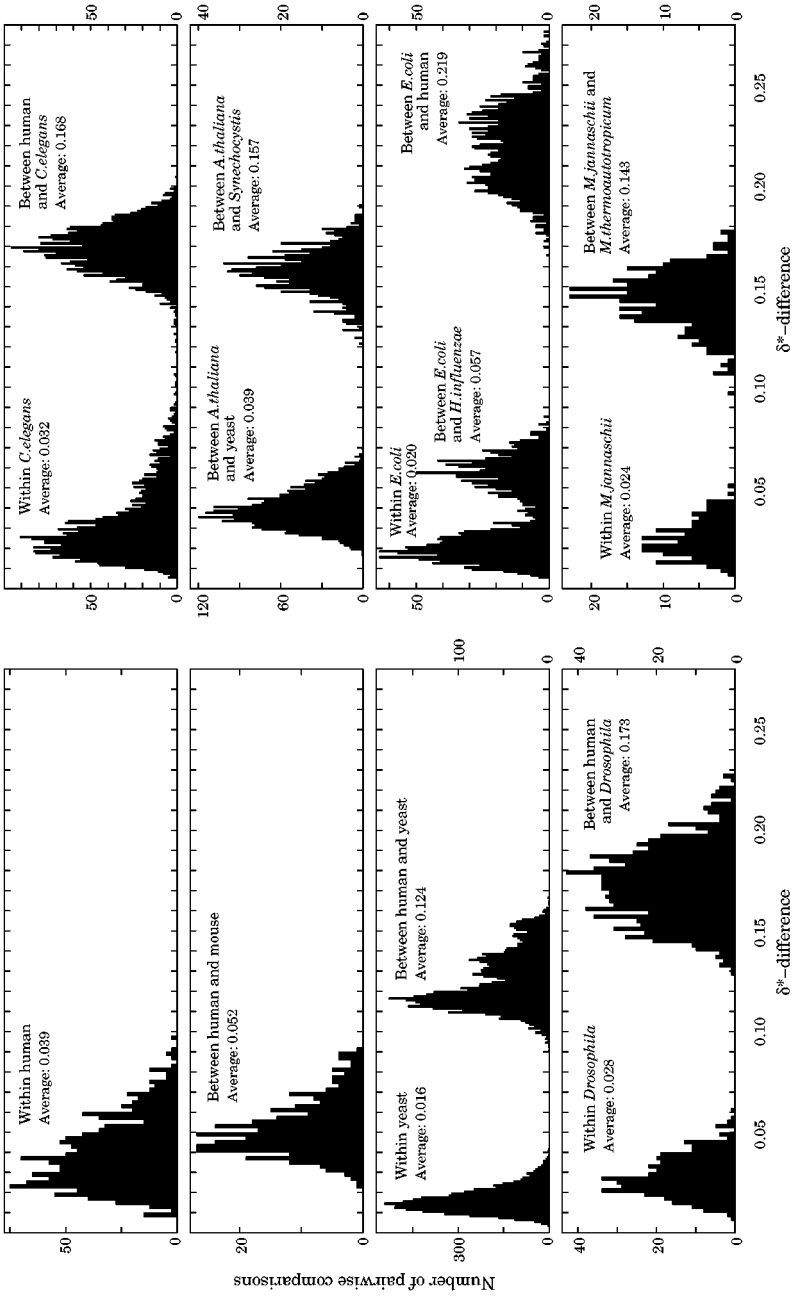


Figure 1 Histograms of δ^*-differences for all pairs of ~50-kb disjoint sequence samples within a single species or from two different genomes.

1. A central unresolved problem concerns whether archaea are monophyletic or polyphyletic. Equivalently, are archaea a separate coherent grouping among prokaryotes? On the basis of rRNA gene comparisons (74, 93), the archaea are deemed monophyletic. This conclusion is supported by some protein comparisons, e.g. the eukaryotic and archaeal RecA-like sequences of Rad51/Dmc1/RadA (14, 82) and the elongation factor EF-1α and EF-2G families (2, 78). However, many protein comparisons challenge the monophyletic character of the archaea. For example, bacterial relationships based on comparisons among the HSP70-kD (*E. coli* DnaK homologue) sequences place the *Halobacteria* closer to the *Streptomyces* than to other archaeal or eukaryotic species (33–35). Further results along these lines apply to the protein families glutamate dehydrogenase (3) and to glutamine synthetase (18). Some of the anomalies are interpreted in terms of lateral transfer events. Lake and collaborators divide the prokaryotes into eubacteria, halobacteria, and eocytes (65, 78).

With respect to genomic signature comparisons, *Sulfolobus* shows the following δ^*-differences to other bacterial genomes: (*Sulfolobus, Clostridium*) $\delta^* = 87$, moderately similar; (*Sulfolobus, Rickettsia/Buchnera*) $\delta^* \sim 130$, distantly similar; (*Sulfolobus*, other thermophilic archaea) $\delta^* \sim 110$–130; (*Sulfolobus*, purple proteobacteria/high G+C Gram$^+$) $\delta^* \sim 190$–270, very distant; (*Sulfolobus*, vertebrates) $\delta^* \sim$ about 90–140, weakly similar. *Halobacterium* δ^*-differences to other prokaryotes are generally very distant ($\delta^* =$ 150–350, mostly >200), excepting δ^* (*Halobacterium, Streptomyces*) $=$ 90–110, weakly similar. The δ^*-differences of *Halobacteria* to the archaeal sequences of *Sulfolobus* sp. and *M. jannaschii* are very distant, $\delta^* > 280$ and >340, respectively. All comparisons with *Sulfolobus* sp. have δ^* values >125 and mostly >180. δ^*-differences of *Halobacteria* from other archaea exceed 245. Thus, a coherent description for the archaea is not supported by the genomic δ^*-difference data. In terms of δ^*-difference, archaea do not behave as a monophyletic clade, and Gram$^+$, Gram$^-$ and archaea tend to be quite diverse clades and intermeshed. The two thermophilic methanobacterial genomes (*M. jannaschii* and *M. thermoautotrophicum*) are distantly similar, $\delta^* =$ 144, and very distant from the halobacterial sequences ($\delta^* = 260$), but weakly similar to the *Sulfolobus* sequences, $\delta^* \sim 120$. These methanogens are very distant from all proteobacterial genomes (generally $\delta^* > 250$) and weakly similar or distant from low G+C Gram$^+$ sequences (δ^* values in the range 110–200).

2. The *Rickettsial* and *Ehrlichial* groups are designated α-proteobacteria on the basis of 16S rRNA. However, these classifications are problematical. The traditional α-types consist of two major subgroups: A_1, including *Rhizobia* and *Agrobacterium tumefaciens*, and A_2, including *Rhodobacter* sp.

and *Paracoccus denitrificans*. A third group, A_3, includes the *Rickettsial* and *Ehrlichial* clades. However, the following global genomic sequence comparisons indicate pronounced discrepancies: (*a*) The A_1 and A_2 genomes are persistently of high G+C content (generally \$60%), whereas A_3 genomes are of low G+C content (<35%). (*b*) The mutual δ^*-differences among A_1 sequences are in the range 45–63 and among the A_2 sequences δ^*-differences register 65–90. The δ^*-differences between A_1 and A_2 traverse the range 62–91. By contrast, the *Rickettsia prowazekii* genome, compared to A_1 and A_2, produces excessive δ^*-differences, generally >200.

3. The mutual δ^*-difference of the two complete *spirochaete* genomes, *B. burgdorferi* and *T. pallidum*, 188, indicates that these genomes are very distant. Moreover, *B. burgdorferi* is very distant from all classical proteobacteria (δ^*-differences mostly >200). *B. burgdorferi* is moderately similar to *Clostridium acetobutylicum* ($\delta^* \sim 87$), weakly similar to a number of other low G+C Gram positive sequences, and δ^* (*B. burgdorferi*, *M. jannaschii*) = 81. In contrast, the *Treponema pallidum* genome is generally moderately to weakly similar to γ- and β-type proteobacterial sequences and to several Gram$^+$ sequences. *T. pallidum* is very distant from the archaeal sequences.

4. The δ^*-differences of *H. pylori* to all other prokaryotic sequences exceed 110 and mostly exceed 160. The sequences weakly similar to *H. pylori* are a few of the γ-proteobacterial sequences and the *B. burgdorferi* genome. However, unlike proteobacterial genomes where the tetranucleotide CTAG is drastically underrepresented, the *H. pylori* genome carries normal representations of CTAG (see below). The *H. pylori* genome sequence has a pathogenicity island about 37 kb in length (*cagA*-region), putatively of "foreign" origin (23). The *cagA*-region is the most deviant in terms of genome signature from the rest of the genome. Specifically, the average δ^*-difference between *cagA* and all other *H. pylori* genomic segments of the same length is 123, significantly higher than δ^*-differences among all other segments (average 31, range 6–110). It appears that δ^*-differences (genomic signature differences) might be used for detecting alien DNA sequences, including pathogenicity islands.

5. *Chlamydia* is very distant from all other eubacteria but remarkably close to *A. fulgidus* ($\delta^* = 47$) and weakly to distantly similar to *P. horikoshii*, *P. aerophilum* and *Sulfolobus* sp. (δ^* values in the interval 100–130). The genome of *A. fulgidus* is moderately to weakly similar to some eukaryotes, *Caenorhabditis elegans*, *Saccharomyces cerevisiae*, and *A. thaliana*, but distant from vertebrates.

BACTERIOPHAGE AND EUBACTERIA For a collection of 23 bacteriophages, it is shown (10) that (*a*) the phage genomes too are endowed with a distinct genome signature; (*b*) the enteric temperate dsDNA phages form a coherent group, in

contrast to the lytic dsDNA phages; and (*c*) the signatures of phages whose replication depends on host machinery converge toward the signatures of the hosts, whereas autologously replicating phages (T4, T7) diverge to their own characteristic signatures. These observations further support the hypothesis (44) that the intrinsic replication and repair mechanisms contribute significantly to the constancy and uniqueness of the species-specific dinucleotide relative abundances.

δ^*-Differences Among Eukaryote Genomes and Between Eukaryotes and Prokaryotes

1. The most homogenous eukaryote genomes occur among fungi (especially *S. cerevisiae*, see Figure 1), whereas the most diverse genomes are found among protists. 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 moderate similarity of human and mouse (Figure 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.

2. The vertebrates show mutual δ^*-differences of moderate similarity. Strikingly, the invertebrates (*D. melanogaster, C. elegans*, and also *B. mori*) are generally distant from vertebrates ($\delta^* > 150$).

3. The dicot *A. thaliana* and *S. cerevisiae* are very close ($\delta^* = 39$).

4. The δ^*-differences of *D. melanogaster* from *E. coli* and *H. influenzae* (both classified as γ-type proteobacteria) are tantalizingly moderately similar. They share the same dinucleotide relative abundance extremes.

ρ^*_{XY}	GC	TA	TT/AA		GC	TA	TT/AA		GC	TA	TT/AA
Dros	+	−	+	; *E.co*	+	−	(+)	; *H. in*	++	−	+
	1.27	0.75	1.24		1.28	0.75	1.22		1.43	0.75	1.25

(Other ρ^*_{XY} are normal.). *T. brucei* is also weakly similar to *D. melanogaster* (49).

5. Most classical eubacteria (e.g. *E. coli, H. influenzae, M. genitalium, M. pneumoniae, Synechocystis* sp.) are very distant from vertebrates, but weakly to distantly similar to *S. cerevisiae* (data not shown). *Methanococcus jannaschii* and *M. thermoautotrophicum* are closer to all eukaryotes than is *P. aerophilum*, again reflecting the very diverse origin and evolution of archaea. Or are archaea generally just deeply divergent prokaryotes that are spread through the eubacterial kingdom?

Mechanisms of the Genome Signature

Mechanisms that underlie the signature determination may include (*a*) context-dependent mutation (of which the methylation/deamination mechanism can be taken as prototypic), or (*b*) selection for structural features of DNA. DNA participates in multiple activities including genome replication, repair, and segregation. In higher eukaryotes, controls on replication can hardly be sequence specific (62). There are fundamental differences in replication characteristics between *Drosophila* and mouse (12). *Drosophila* DNA replicates frenetically in the first hour after fertilization, with replication bubbles distributed about every 10 kb. At about 12 h, effective origins are spread to about 40 kb apart. In mouse, the rate of replication appears to be uniform throughout developmental and adult stages. Cell divisions involve DNA stacking on itself and loopouts that need to be judiciously decondensed to undergo segregation. The observed narrow limits to intragenomic heterogeneity may correlate with conserved features of DNA structure.

The influence of the (double-stranded dinucleotide) base step on DNA conformational preferences is reflected in slide, roll, propeller twist, and helical twist parameters (21, 39). Calculations and experiments both indicate that the sugar-phosphate backbones are relatively flexible. 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. Inappropriate juxtaposition or distribution of purine and pyrimidine bases could engender steric clashes (39). 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 have fewer steric conflicts between neighbors (37, 39). Dinucleotide relative abundance deviations may reflect duplex curvature, supercoiling, and other higher-order DNA structural features. Many DNA repair enzymes recognize shapes or lesions in DNA structures more than specific sequences (26, 63). Nucleosome positioning, interactions with DNA-binding proteins, and ribosomal binding of mRNA appear to be strongly affected by dinucleotide arrangements (21, 91).

Other general influences relate to environmental conditions affecting DNA sequence and structure include osmolarity gradients, UV irradiation, temperature extremes, hydrostatic pressures, pH environment, metal concentrations, habitat variants, energy sources and systems, interacting fauna and flora, and stress conditions that can trigger transposition events and alternative recombination pathways. Further factors that affect genomic structure and organization and flux of DNA involve direct or indirect transfer of genomic pieces between organisms.

FREQUENT AND RARE WORDS (OLIGONUCLEOTIDES) IN SOME PROKARYOTE GENOMES

It is of interest to determine which words of moderate size in the genome occur with unusually high or low frequencies and to identify anomalies in their distribution. For DNA, rare words might be binding sites for transcription control factors restricted to specific locations. Alternatively, rare words may be discriminated against due to structural defects (kinking), e.g. as has been suggested for the tetranucleotide CTAG, which is extremely rare in most purple proteobacterial genomes (20). The crystallographic resolution of the TrpR-DNA complex (75) and also for the MetJ-DNA complex (76) indicates CTAG kinks that may be structurally deleterious elsewhere in the DNA. The potential role of the *vsr* gene product (very short patch repair system) in attenuating the frequency of CTAG in certain bacterial genomes is also recognized (7, 45).

Frequent words often include parts of repetitive structural, regulatory and transposable elements, e.g. uptake signal sequences in *H. influenzae* (87) and Chi sites of *E. coli* (which in association with the RecBCD complex promote recombination). [For the formal statistical theory of rare and frequent words, see (47, 50, 55)]. In proteins, frequent oligopeptides often reflect characteristic motifs shared in certain protein families, e.g. the sequence environment of the catalytic triad of serine proteases, the ATP-binding motif (Walker-box) of prokaryotic and eukaryotic proteins. A comparison of texts or distributions of such words within sets of sequences from different organisms may suggest important evolutionary tendencies or constraints at work.

A remarkably frequent word called the Highly Iterated Palindrome GGCGATCGCC (see 79) occurs in the cyanobacterium, *Synechocystis* sp. (PCC 6803), genome (2768 occurrences). The principal frequent words of *M. genitalium* are related to multiple long trinucleotide iterations of (GTA), (CTT), and (CTA).

In *H. influenzae*, three major classes of frequent oligonucleotides stand out: (*a*) oligonucleotides related to uptake signal sequences (USSs), AAGTGCGGT (USS^+) and its inverted complement (USS^-); (*b*) multiple tetranucleotide iterations (e.g. $(CCAA)_{37}$, $(CCAA)_{21}$, $(TCAA)_{33}$, $(TCAA)_{23}$), and others; (*c*) Intergenic Dyad Sequences (IDSs) found as AAGCCCACCCTAC and its dyad form (71). The USS^+ and USS^- occur in almost equal counts that are remarkably evenly spaced around the genome and that appear predominantly in the same reading frame in protein coding domains (USS^+ translated to Ser-Ala-Val, USS^- translated to Thr-Ala-Leu). These observations suggest that USSs contribute to global nonspecific genomic functions, for example, in replication and/or repair processes, or as membrane attachments sites, or as sequences helping

to pack DNA. The extensive tetranucleotide iterations (i.e. unknown in prokaryotes other than *H. influenzae*), through polymerase slippage during replication and/or homologous recombination may produce subpopulations expressing alternative proteins. The 13-bp frequent IDS words, AAGCCCACCCTAC and its inverted complement, invariably intergenic, occur mostly in clusters and provide potential for various secondary structures, suggesting that these sequences may be important signals for regulating the activity of flanking genes (71).

In *Neisseria gonorrhoeae*, constitutive natural uptake of DNA of its own genus is related to the oligonucleotides TTCAGACGGC and its inverted complement GCCGTCTGAA, which are the most frequent words of size 10 in *N. gonorrhoeae* DNA. By contrast, the *Bacillus subtilis* genome contains no frequent oligonucleotides.

The most notable frequent words of *M. jannaschii* are parts of the 30-bp oligonucleotide W = RTTAAAATCAGACCGTTTCGGAATGGAAAY (R = purine, Y = pyrimidine), with 63 occurrences and 3 in its inverted complementary form. Allowing for $\geq$80% identity, 134 such words occur in the genome. These words mostly occur in clusters separated by 5 long gaps of 130–400-kb lengths. Within the clusters, the words tend to be regularly spaced and separated by approximately 40 bp. These words constitute "short repeat segments" of a multicopy repeat structure (19).

The frequent word analysis applied to the genome of *Methanobacterium thermoautotrophicum* (1.75 Mb) (86) revealed 124 perfect copies of the 30-bp oligonucleotide W^* = ATTTCAATCCCATTTTGGTCTGATTTTAAC and 47 copies of its inverted complement, with no other occurrences allowing up to 6 errors. All 124 occurrences of W^* are clustered in the 8-kb region 983325–991536 and all 47 occurrences of the inverted complement are in the 3-kb region 1472410–1475423. Spacings between these words in the clusters range from 64 bp to 80 bp, with insignificant similarity. W and the inverted complement of W^* mismatch only at seven positions.

The *Archaeglobus fulgidus* genome (59) contains 60 copies of the 30-bp oligonucleotide W^{**} = CTTTCAATCCCATTTTGGTCTGATTTCAAC. All copies of W^{**} are confined to the 4-kb region 2089294 to 2093359. There are no variants of W^{**} in the *A. fulgidus* genome, allowing up to 6 mismatch errors. Notably, there are 47 exact occurrences of the inverted complement to W^{**} and one occurrence with one mismatch error and no others with at most 6 mismatched errors. The inverted complement words cluster between positions 1691936 to 1695157 (about 3.2 kb) displaced about 1/2 Mb from the other cluster.

The archaeon *Pyrobaculum aerophilum* genome contains 76 precise copies of the 24 nucleotide word V = CTTTCAATCCTCTTTTTGAGATTC all in

a single cluster of ~5 kb in length, and 3 additional copies (showing up to 4 errors) in the same cluster. There are no copies (accommodating up to 4 errors) on the complementary strand. The first 15 nucleotides of V and W^* differ at only 3 positions.

The current GenBank DNA data base, totaling in excess of 700 Mb, was screened for occurrences of W, W^*, and W^{**}, allowing up to 6 mismatches. Strikingly, only three occurrences were detected, each with six errors, one among *C. elegans* sequences and two among mouse sequences. None of the W, W^*, or W^{**} was found in classical eubacterial genomes presently available.

The archaeal *Pyrococcus horikoshii* genome (1-Mb contig, available from National Institute of Technology and Evaluation, Japan, via www (http://www.nite.go.jp/)] contains the 29-bp oligonucleotide U = CTTTCCACACACTATTTAGTTCTACGGAAAC at 69 places and 2 exact occurrences of its inverted complement U' distributed to three clusters. The first cluster includes 18 occurrences about evenly spaced traversing the region 65633–66742 (about 40 bp separating successive occurrences).

Allowing up to 6 errors, U' increases to 26 occurrences (predominantly GTTTCCGTAGAACTcAgTAGTTGGAAAG) confined to 183079–184834 about evenly spaced. The third cluster of 66 copies extends from 966566–970971, again evenly spaced with about 40 bp separating each pair of U. There is no unambiguous similarity between U, V, and W. Corresponding repeats were not found in any nonarchaeal genomes. The significance of these repeats is unknown.

Distributional Properties of Some Frequent Oligonucleotides

We describe several distributional anomalies of the USS sequences of *H. influenzae* analyzed with the assistance of *r-scan* statistics [for background and applications of *r-scans*, see (13, 15, 42, 43, 55)].

OVERDISPERSIONS AND CLUSTERS APPLIED TO THE COMBINED SET OF USS$^+$ AND USS$^-$ OCCURRENCES Significant overdispersion is detected at positions 1.56–1.59 Mb, a region dominated with phage Mu-like sequences. A second significant overdispersion of USSs occurs in the region of positions 834–855 kb, which is replete with ribosomal protein genes. A significant cluster is found at 1.756 Mb associated with a 168-bp coding sequence (containing a USS dyad) tandemly repeated four times.

SIGNIFICANTLY EVEN SPACINGS OF USS IN EACH ORIENTATION Another striking anomaly of USS positions concerns the significantly even spacings of the USS$^+$ occurrences and the same for the USS$^-$ occurrences. Specifically, both USS$^+$ positions and USS$^-$ positions have respective minimum spacings significantly higher than expected by chance, with the probability <0.001 to

observe such an even distribution with the same numbers of randomly distributed markers.

Comparable to the foregoing, the *r-scan* lengths ($r = 1, 2, \ldots, 6$) revealed an excessively even distribution of the highly iterated palindrome HIP1 GGCGATCGCC in the *Synechocystis* sp. genome. The even spacing of HIP1 ($p^* \ll 0.1\%$) is more extreme than that of USSs in *H. influenzae*. The critical minimum spacing for 0.1% significance is 9 bp, i.e. the chance that all spacings are >9 bp has probability <0.001 for a random distribution of HIP1 words. The observed minimum *r-scan* is 52 bp.

CTAG Underrepresentations

CTAG is significantly underrepresented in many bacteria encompassing purple proteobacteria (exceptions *H. pylori* and *N. meningitidis*), high-G+C Gram-positive *Streptomyces*, and several archaeal genomes but generally not in eukaryotes. Although the tetranucleotide CTAG is very low in *E. coli* and *H. influenzae* (Table 4), the distribution of CTAG sites around the *E. coli* genome shows six significant clusters each contained in a rRNA unit (45), whereas in the *H. influenzae* genome, the *r*-scan statistics (55) demonstrate that the extant CTAG sites are randomly distributed. The relative clustering of seven to nine CTAG sites in every *E. coli* rRNA gene about once every 400 bp is in sharp contrast to the mean frequency of CTAG in *E. coli* of about one per 5200 bp over the whole genome. This anomaly applies to numerous other proteobacterial genomes. CTAG is generally low in most classes of *E. coli* phages (10). Exceptions are P4 and Mu ($\tau^* = 0.93$ and 0.97, respectively). The CTAG sites tend to occur in small clusters in each of these phages.

Agrobacterium tumefaciens is significantly low in CTAG ($\tau^* = 0.65$), whereas its associated Ti plasmid sequence (106 kb) possesses $\tau^*_{CTAG} = 0.86$ in the normal range (data not shown). *N. gonorrhoeae* is normal for CTAG but is severely underrepresented for CATG and GATC. Except for *Streptomyces* genomes (e.g. *S. griseus*, *S. lividans*, and *S. coelicolor* [$\tau^* < 0.50$]), CTAG is normally representated in most other Gram-positive sequence sets, including all low-G+C Gram-positive types, together with the high-G+C Gram-positive sequences of *M. tuberculosis*, and *M. leprae*. Moreover, CTAG is normally representated in all cyanobacterium sequences ($0.94 < \tau^*_{CTAG} < 1.04$) and is in the low-to-normal range for all mycoplasmas (*M. genitalium*, $\tau^*_{CTAG} = 0.95$; *M. capricolum*, $\tau^*_{CTAG} = 0.83$) and low normal in *Borrelia burgdorferi*.

Archaeal sequences vary in CTAG occurrences. Whereas the methanothermophiles, including *M. thermoautotrophicum* and *M. jannaschii*, are significantly low, *P. aerophilum* and *Sulfolobus* sp. have CTAG relative abundances in the normal range (56). The *M. jannaschii* genome is unsurpassed in the extremely low relative abundance value of its CTAG tetranucleotides. Specifically,

Table 4 Tetranucleotide extremes (τ^* values)[a] for genomes with large aggregate DNA sequences available

	Palindromic										Other
	CTAG	GTAC	CATG	GATC	TCGA	ACGT	CCGG	GGCC	CGCG	GCGC	TATA	
Escherichia coli	--- 0.24											CAAG/CTTG – (0.73)
Haemophilus influenzae	-- 0.63		--- 0.43		– 0.78		--- 0.37	--- 0.50	-- 0.70	-- 0.62	– 0.71	CCGC/GCGG + (1.25)
Neisseria gonorrhoeae	-- 0.64		-- 0.61	--- 0.46				-- 0.57				
Neisseria meningitidis			-- 0.62	--- 0.44				-- 0.51				CTAA/TTAG – (0.76)
Helicobacter pylori		--- 0.16			--- 0.12	--- 0.16	++ 1.43	-- 0.67		-- 0.68		GACC/GGTC CCTC/GAGG GCGA/TCGC – (0.74) – (0.74) ++ (1.42)
Bacillus subtilis	(no significant tetranucleotide extremes in *B. subtilis*)											
Streptococcus pyogenes				– 0.74			– 0.73					
Clostridium acetobutylicum				– 0.76				– 0.75				
Mycobacterium leprae	(no significant tetranucleotide extremes in *M. leprae*)											
Mycobacterium tuberculosis	(no significant tetranucleotide extremes in *M. tuberculosis*)											
Mycoplasma genitalium											– 0.78	CCGA/TCGG – (0.76)
Mycoplasma pneumoniae											– 0.78	
Synechocystis sp.									--- 0.37	-- 0.63		
Deinococcus radiodurans	--- 0.37			-- 0.64								CTAA/TTAG CTAC/GTAG AACT/TTGA ATAA/TTAT -- (0.70) ++ (1.39) ++ (1.37) + (1.23)
Treponema pallidum	(no significant tetranucleotide extremes in *T. pallidum*)											
Borrelia burgdorferi	(no significant tetranucleotide extremes in *B. burgdorferi*)											
Chlamydia trachomatis	(no significant tetranucleotide extremes in *C. trachomatis*)											
Methanococcus jannaschii	--- 0.06	--- 0.31		--- 0.11				--- 0.43		--- 0.32		CAGG/CCTG – (0.77)
Methanobacterium thermoautotrophicum	--- 0.33											CTAA/TTAG ATAG/CTAT -- (0.68) + (1.29)
Archaeoglobus fulgidus	--- 0.22			--- 0.45					-- 0.54	-- 0.67		
Pyrococcus horikoshii								++ 1.34		-- 0.62		GGAC/GTCC GACC/GGTC CGGC/GCCG – (0.71) – (0.78) + (1.28)
Pyrobaculum aerophilum												CAGG/CCTG – (0.77)

[a]The relative abundance of the tetranucleotide XYZW is defined as $\tau^*_{XYZW} = (f^*_{XYZ} f^*_{XZW} f^*_{XYW} f^*_{YZW} f^*_X f^*_Y f^*_Z f^*_W)/(f^*_{XY} f^*_{YZ} f^*_{ZW} f^*_{XZ} f^*_{YW} f^*_{XW})$, where f^* are symmetrized frequencies of the indicated mono-, di- or trinucleotides (dots stand for gaps; see Karlin and Cardon 1994). Only those tetranucleotides are reported that possess both τ^* value and Markov abundance ($f^*_{XYZ} f^*_{XYZ}/f^*_{YZ}$) in low/high range. No tetranucleotide extremes occur in human, *D. melanogaster, C. elegans*, yeast, and *A. thaliana* sequences.

over the *M. jannaschii* 1.66-Mb genome, there are only 90 CTAG sites, yielding the very low relative abundance value $\tau^* = 0.06$. Their distribution is highly anomalous, exhibiting two major clusters and several significantly large gaps. For example, 9 CTAG sites occur in the region from 154904 to 160584, and 10 counts of CTAG occur in the region from 636994 to 643016. CTAG in *M. thermoautotrophicum* is about as low as in *E. coli*. Also, their spacings around the genomes are highly anomalous. An *r-scan* statistical (56) analysis of their distribution reveals four clusters in the region of positions 41655–42267, in 51738–52607, in 1605403–1606469, and in 17217128–1723045. Intriguingly, the latter two clusters overlap the two rRNA operons of *M. thermoautotrophicum*, the first located in the 6-kb stretch 1607572–1609150 and the second located in the region 1717850–1724357. Are CTAG sites possible binding sites for regulatory proteins and/or possible nucleation sites in the formation of ribosomes?

Other Tetranucleotide Extremes

The palindromic tetranucleotides CCGG and GGCC of *H. influenzae* have markedly low representations, and these sites tend to be clustered about rRNA sequences (55). The same bias and distribution apply to CTAG sites in *E. coli*.

Tetranucleotide biases in eukaryotes are relatively uncommon; all genomes with substantial DNA available show no significant tetranucleotide over- or underrepresentations. Most underrepresented tetranucleotides occur in prokaryotes. *M. jannaschii* is very significantly low in five palindromic tetranucleotides, whereas *M. thermoautotrophicum* only is underrepresented in CTAG. *M. genitalium* and *M. pneumoniae* show the identical low extreme for TATA. The two spirochaetes *B. burgdorferi* and *Treponema pallidum* carry no tetranucleotide extremes. The same applies to *M. leprae* and *M. tuberculosis*.

H. influenzae is distinguished with eight low palindrome tetranucleotides. *H. pylori* is uniquely overrepresented for CCGG, and *P. aerophilum* is uniquely overrepresented for GGCC.

Restriction Avoidance

The low values for palindromic tetranucleotides in Table 4 may reflect to some extent restriction avoidance by the various prokaryotes. The *M. jannaschii* genome (1.66 Mb complete) features five significantly low palindromic tetranucleotides and one high nonpalindromic tetranucleotide. On the basis of sequence similarity, eight potential methylases of restriction modification systems have been reported (R Roberts, personal communication). The counts and distributions of the palindromic nucleotides {CTAG, GATC, GTAC, CATG} of the same nucleotide content are striking. For example, CTAG occurrences are drastically low, confined mainly to two significant clusters about kilobase positions 155 to 161 and 637 to 643, the latter cluster intercalated with seven

putative tRNA genes. GATC sites tally 252 counts distributed in five significant clusters about kilobase positions 158 to 159, 349 to 352, 530 to 532, 638 to 640, and 664 to 673, two of which coincide with the CTAG clusters. There are three significantly long gaps of 70, 71, and 117 kb devoid of GATC sites (*r-scan* statistics). GTAC counts are 334, highlighting again the same two clusters at kb 155 to 159 and 639 to 643. In sharp contrast, CATG sites show a normal count of 3554 occurrences, quite randomly distributed around the genome.

GCGC and CGCG tally 119 and 101 counts, respectively, in *M. jannaschii* distributed around the genome featuring clusters in the same regions, about positions 155 to 161 and 637 to 643. A propos, a profile of G+C counts in 10-kb windows (or 50-kb windows) highlights two regions concentrated about positions 155 to 161 and 637 to 643 with G+C frequencies near 50%, contrasted to a global genome of 31% G+C content.

CODON BIASES IN BACTERIAL GENOMES

The nature of codon choices varies considerably from organism to organism [for a recent review, see (85)]. Our objective in this part is to highlight some new perspectives and results on codon biases in selected complete genomes.

Variations in tRNA availabilities are interpreted by several authors as a key factor in producing codon bias of the "highly expressed genes" of yeast and *E. coli*. Translational accuracy and efficiency and codon/anticodon interaction strength are also influential (1, 64). Selective and nonselective substitutional biases operating during DNA replication, transcription, and repair processes also play a role. Compartmental heterogeneity (isochores) in mammalian genomes underscore S = (C+G) or W = (A+T) nucleotide predominance (38). Other factors that may influence codon choices in vertebrates include CpG suppression, methylation effects of DNA (90), tissue or organ specificity (38), mRNA stability (1), codon context (52, 57), and species of origin (66).

Establishing the rules of codon usage is of interest with respect to fundamental evolutionary questions. Some preliminary analysis suggests that recently imported genes show deviant codon usage from the host gene inventories (66, 67, 70). A deeper understanding of codon and residue choices can help in gene prediction, in characterizing properties of a given gene and in classifying gene families.

Comparisons of Codon Usage Between Different Gene Classes

Variation in codon usage across a genome can be assessed in many ways. One approach is to compare codon usage within and between various gene classes of the organism. For example, the genes of bacterial genomes have been divided

into 14 major function and cellular classes [adapted from (77)], each generally comprised of several subclasses. Another means in forming gene classes can be based on partitioning the genome into 100-kb, 200-kb, or longer contigs and assembling all genes of each contig to define a gene group (S Karlin & J Mrázek, unpublished).

Gene groups can be generated by forming k (e.g. $k = 2, 3, 5, 10$) clusters distinguishing genes by similarity of codon usage (in 61 dimensional space) (70), or alternatively by similarity of amino acid usages or relative to a reduced set of amino acids or codons. The different clusters can be regarded as distinct gene classes.

Measures of Relative Codon Biases

CODON ADAPTATION INDEX A quantitative measure proposed for assessment of codon bias is the codon adaptation index [CAI, (84)]. This specifies a reference set of genes, almost invariably, $\mathcal{H}$, chosen from among "highly expressed genes." Defining $w_{xyz} = f^{\mathcal{H}}_{xyz} / \max_{xyz \varepsilon a} f^{\mathcal{H}}_{xyz}$ as the ratio of the frequency of the codon (xyz) to the maximal codon frequency in $\mathcal{H}$ for the same amino acid a, the CAI of a gene of length L is taken as $(\prod_{i=1}^{L} w_i)^{1/L}$ (the log average), where i refers to the i^{th} codon of the gene and w is calculated as above. High values (near 1) of CAI correlate with high expression levels. Classification of genes according to their CAI values has been done in several publications. Genes that are known (experimentally) to be highly expressed, at least during cellular fast growth, include most ribosomal protein genes and genes coding for elongation factors (tuf and fus) and some membrane genes. However, not all ribosomal proteins have a high CAI value (57).

CODON BIAS (CB) BETWEEN GENE CLASSES We introduce a flexible way to assess bias of one group of genes (or a single gene) relative to a second group of genes (57). Let $\mathcal{C}$ be a class of genes with aggregate codon frequencies $c(x, y, z)$ normalized to 1 relative to each amino acid so that $\sum_{(x,y,z)=a} c(x, y, z) = 1$, where the sum extends over all codons (x, y, z) translated to amino acid a. Let $\{f(x, y, z)\}$ indicate the codon frequencies for the gene family F, also normalized to 1 in each codon family. We assess the codon bias of the gene family F relative to the gene family $\mathcal{C}$ by the formula

$$B(\mathcal{F}\,|\,\mathcal{C}) = \sum_{a} p_a(\mathcal{F}) \sum_{(x,y,z)=a} |f(x, y, z) - c(x, y, z)| \qquad [1.]$$

where $\{p_a(\mathcal{F})\}$ is the set of amino acid frequencies of the combined genes of F. Notice the asymmetry of $B(\mathcal{F}\,|\,\mathcal{C})$ in that only the amino acid frequencies of F appear as weights. We refer to the gene collection $\mathcal{C}$ as the standard to which different gene groups $\mathcal{F}^{(1)}, \mathcal{F}^{(2)}, \ldots, \mathcal{F}^{(r)}$ are compared. The formula [1.] can

also be applied to a subset of amino acids (e.g. restricted to charge or aromatic amino acids). Some preliminary results for calculation of codon biases over different gene classes are outlined next.

Anomalies of Ribosomal Proteins

The ribosomal protein family codon frequencies generally deviate strongly from overall codon frequencies in many bacterial genomes (Table 5). The greatest disparity occurs for the *E. coli* and *B. subtilis* genomes about the same magnitude of difference, $B(\mathrm{RP} \mid G_{E.\ coli}) = 0.520$ and $B(\mathrm{RP} \mid G_{B.\ sub.}) = 0.567$. This strong

Table 5 Relative codon bias[a] for three gene collections (all genes, ribosomal proteins, and amino acyl tRNA synthetases) in several complete baterial genomes

	E. coli			*H. influenzae*			*H. pylori*			*B. subtilis*		
$\{\mathcal{F}\}\backslash\{C\}$	**all**	**RP**	**tRN**	**all**	**RP**	**tRN**	**all**	**RP**	**tRN**	**all**	**RP**	**tRN**
all	*	530	297	*	398	179	*	107	70	*	555	163
RP	520	*	289	399	*	317	114	*	156	567	*	475
tRN	284	271	*	177	297	*	67	145	*	159	460	*
# of genes	**4283**	**55**	**22**	**1680**	**50**	**21**	**1578**	**52**	**21**	**4098**	**52**	**24**

	M. genitalium			*M. pneumoniae*			*Synechocystis sp.*			*B. burgdorferi*		
$\{\mathcal{F}\}\backslash\{C\}$	**all**	**RP**	**tRN**	**all**	**RP**	**tRN**	**all**	**RP**	**tRN**	**all**	**RP**	**tRN**
all	*	132	65	*	134	100	*	223	91	*	157	47
RP	135	*	181	137	*	176	223	*	238	160	*	183
tRN	64	172	*	96	174	*	90	232	*	47	175	*
# of genes	**466**	**50**	**20**	**677**	**50**	**19**	**3168**	**53**	**22**	**851**	**53**	**20**

	M. jannaschii			*M. thermoautotrophicum*			*A.fulgidus*		
$\{\mathcal{F}\}\backslash\{C\}$	**all**	**RP**	**tRN**	**all**	**RP**	**tRN**	**all**	**RP**	**tRN**
all	*	270	101	*	160	93	*	121	117
RP	256	*	196	161	*	177	126	*	105
tRN	100	211	*	92	178	*	116	96	*
# of genes	**1686**	**60**	**18**	**1869**	**61**	**17**	**2408**	**61**	**19**

[a]See formula [1]. All values are multiplied by 1000.

codon bias of RP genes holds also for the *H. influenzae*, *Synechocystis*, and *M. jannaschii* genomes. By contrast, codon usage of the ribosomal proteins in the two reported Mycoplasma genomes (*M. genitalium*, and *M. pneumoniae*), *H. pylori*, and *A. fulgidus* is largely similar to that of the average gene. The aminoacyl tRNA synthetases (tRN) have codon frequencies more similar to the average gene (all) compared to RP, generally by a factor of two or more. The foregoing results are consistent with the proposition that genes highly expressed during exponential growth phase, which certainly include ribosomal proteins, show highly biased codon usages. However, tRNA synthetases are also essential genes and putatively highly expressed in the same environment, but the codon bias is much reduced.

Why do the ribosomal proteins often register the largest codon bias in *E. coli* and *B. subtilis* with respect to their genomes, but markedly less for the other complete genomes of *H. influenzae*, *M. genitalium*, and *M. jannaschii*? This may, in part, be due to the fast-growing nature of *E. coli* and *B. subtilis* compared to other prokaryotes.

YEAST The yeast (*S. cerevisiae*) (Table 6) nuclear RP codon usages are extravagantly deviant from the average yeast protein, B(RP-nuclear yeast | all-yeast) = 0.743, whereas the mitochondrial RP codon frequencies of yeast are modestly similar to the average nuclear gene codon frequencies B(RP-mitochondrial | all-yeast) = 0.163. The tRN nuclear and mitochondrial genes produce moderately similar biases, namely B(tRN-nuclear | G-yeast) = 0.175, B(tRN-mitoch. | all-yeast) = 0.107.

Table 6 Relative codon bias[a] for five gene collections (all genes, nuclear ribosomal proteins, mitochondrial ribosomal proteins, nuclear aminoacyl tRNA synthetases, and mitochondrial amino acyl tRNA synthetases) in complete yeast genome

{F}\{C}	all	nuc RP	mt RP	nuc tRN	mt tRN
all	*	730	170	179	107
nuc-RP	743	*	781	563	844
mt-RP	163	763	*	275	184
nuc-tRNA	175	547	268	*	266
mt-tRNA	107	809	190	267	*
# of genes	6067	53	11	12	6

[a]See formula [1]. All values are multiplied by 1000.

Relative Codon Usage Variation Among Bacterial and Yeast Genomes

The average difference of codon usage of each genome relative to the other genomes generally exceeds 300 (Table 7). The closest to *E. coli* is *B. subtilis* with $B(Bsu \mid Eco) = 274$ (see legend to Table 7 for species abbreviations). Codon usage substantially deviant from *E. coli* genes occurs for the genes of *H. influenzae*, ($B(Hin \mid Eco) = 518$), *M. genitalium*, ($B(Mge \mid Eco) = 615$), *M. jannaschii*, ($B(Mja \mid Eco) = 677$) and *M. thermoautotrophicum* ($B(Mth \mid Eco) = 605$). The greatest codon biases relative to *H. influenzae* are seen for the genes of *Mth* and *Afu*; the least occurs for *Mge*. *Hpy* and *Mth* differ significantly in codon frequencies. *B. subtilis* as a standard entails codon bias <500 from all other genomes (except *Mth*). *M. thermoautotrophicum* (*Mth*) genes show codon biases persistently extreme relative to the other bacterial genomes (excepting *A. fulgidus*), all in excess of 530 (mostly >600 and several >700). Notably, $B(Afu \mid Mth)$ is only 279. By contrast, codon bias of *M.*

Table 7 Relative codon bias[a] among complete bacterial and yeast genomes (multiplied by 1000)

$\{\mathcal{F}\}\backslash\{C\}$	Eco	Hin	Hpy	Bsu	Mge	Mpn	Syn	Bbu	Mja	Mth	Afu	Sce	G+C
Eco	*	522	440	293	673	395	354	741	766	589	521	473	51%
Hin	518	*	369	402	289	365	416	348	445	802	770	369	38%
Hpy	400	355	*	318	393	357	309	421	482	762	653	354	39%
Bsu	274	394	327	*	462	380	311	477	495	541	459	303	44%
Mge	615	276	382	442	*	393	462	233	286	834	786	335	32%
Mpn	398	379	355	385	402	*	276	520	524	616	559	278	40%
Syn	363	435	339	335	540	280	*	609	627	631	612	367	48%
Bbu	661	328	420	465	244	522	538	*	220	821	778	392	29%
Mja	677	444	490	477	299	539	562	214	*	749	720	380	31%
Mth	605	780	718	536	787	606	616	764	756	*	289	558	50%
Afu	552	768	625	466	758	565	616	735	720	279	*	507	49%
Sce	443	363	355	300	331	291	346	373	377	577	532	*	38%

Species are abbreviated as follows: *Escherichia coli* (Eco, includes 4283 annotated genes and ORFs), *Haemophilus influenzae* (Hin, 1680 genes), *Helicobacter pylori* (Hpy, 1578), *Bacillus subtilis* (Bsu, 4098), *Mycoplasma genitalium* (Mge, 466), *Mycoplasma pneumoniae* (Mpn, 677), *Synechocystis* sp. (Syn, 3168), *Borrelia burgdorferi* (Bbu, 851), *Methanococcus jannaschii* (Mja, 1680), *Methanobacterium thermoautotrophicum* (Mth, 1869), *Archaeoglobus fulgidus* (Afu, 2408) and *Saccharomyces cerevisiae* (Sce, 6067).

[a]See formula [1]. All values are multiplied by 1000.

jannaschii standard versus *M. genitalium* is strikingly low, 286, and otherwise mainly >440.

Site 3 G+C Frequencies Around the Genome

Each of the bacterial genomes were partitioned into ten contigs of about equal lengths. The genes of each contig were assembled into a gene class. Figure 2 depicts the variation of site 3 G+C frequencies for these ten gene classes. *E. coli* and *B. subtilis* in the ter contig show S3% reduced by at least 5% from genes near ori-C. The *B. subtilis* S3% value is "symmetric" about ori-C or ter increasing to a maximum about halfway between ori-C and the ter region in both halves. S3 frequencies in *H. influenzae* increase slightly in both directions from ori-C to a maximum in the ter contig. The archaeal *M. jannaschii* and *A. fulgidus* S3 frequencies are constant around their genomes, whereas *M. thermoautotrophicum* is manifestly variable. Synechocystis is also rather constant. These results support speculations connecting replication timing to codon usage and to the possibility of multiple replication origins in several of these genomes.

Codon Bias and "Alien" Genes

Genes within a species tend to be rather homogeneous in base composition and in amino acid and codon usages, although the "highly expressed genes" in bacterial genomes during exponential growth phase are often significantly different in codon usage and to a lesser extent in amino acid usage from the average gene. Prototypes of highly expressed genes in bacterial genomes include ribosomal proteins, translation elongation factors, major chaperonins and some outer membrane proteins. Other genes with high codon bias may be considered to be DNA imported through recent horizontal transfer or to be deviant due to other disrupting influences. In terms of our codon bias assessments, we characterize genes as "alien" if they fulfill the following criteria: (*a*) codon bias (formula [1]) of gene *g* compared to the average gene of the species exceeds an appropriately high threshold; (*b*) codon bias of *g* relative to the set of ribosomal proteins B ($g \mid \mathrm{RP}$) is also appropriately high. Requirement (*b*) excludes most "highly expressed genes" as alien genes. At the time of introgression, horizontally transferred genes reflect the genome composition of the donor genome that, however, over time shift to the DNA compositional "biases and asymmetries" of the new genome (66, 67).

For the *B. subtilis* genome, Figure 3 plots the codon biases of all long individual genes (those of length at least 200 codons) relative to the average gene on the vertical axis and to the class of RP genes on the horizontal axis [see (57) for the corresponding analysis of *E. coli* genes]. Alien genes are defined such that $\mathrm{B}(g \mid \mathrm{all}) > 0.42$ and $\mathrm{B}(g \mid \mathrm{RP}) > 0.45$. By these criteria, we distinguish

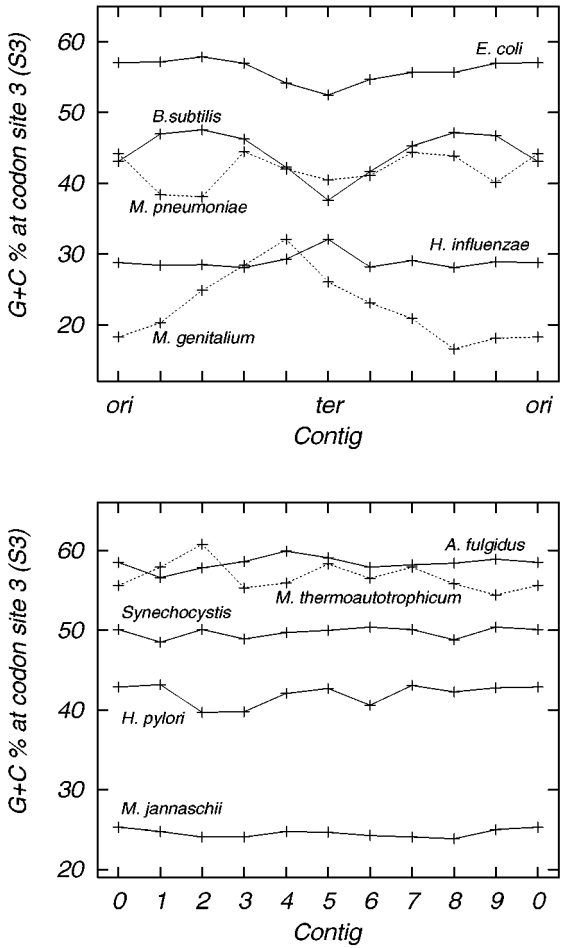


Figure 2 S3 (G+C at codon site 3) variation along the complete prokaryotic genomes. Each genome was divided into ten disjoint contigs of equal lengths. For genomes with known localization of the origin of replication (*upper panel*), the first contig (shown both at left and at right) is centered at the origin of replication. Opposite to the origin of replication is the contig containing the ter-region (in the middle of the plot). For genomes where a unique origin of replication was not identified (*lower panel*), the position of the first contig is arbitrary.

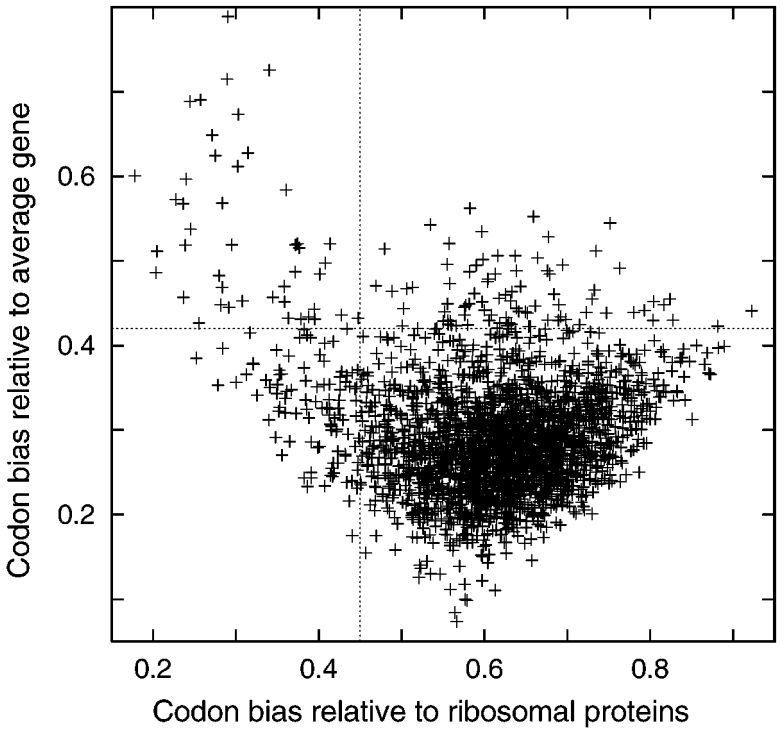


Figure 3 Each *B. subtilis* gene of ≥200 codons is represented by a point with coordinates corresponding to its codon bias relative to the average gene and codon bias relative to ribosomal proteins. Thresholds for identifying alien and highly expressed genes are indicated by dashed lines.

88 alien genes, including 77 ORFs of unknown function. The distribution of alien genes contains seven clusters C1, C2, . . . , C7 consisting almost entirely of ORFs: C1 contains 6 or 7 genes in a 10-kb segment (first gene of C1 starts at position 546697—last gene of C1 starts at position 556476); C2, 2 genes (654940–656245); C3, 4 or 5 genes (737109–744359); C4, 2 genes (2068284–2070341); C5, 3 genes (3463825–3466120); C6, 8 genes (4124150–4137998); and C7, 4 genes (4172873–4175077).

The highly expressed genes are defined by the codon bias values $B(g \mid all) > 0.42$ but $B(g \mid RP) < 0.45$. Table 8 lists all long genes (>200 codons) satisfying these criteria. These include 6 large RPs; the elongation factors EF-G, EF-Tu, EF-Ts; a number of mainstream glycolysis genes (triose phosphate isomerase, phosphoglycerate kinase, g3pd, enolase, aldolase, pyruvate dehydrogenase E1, E2 and E3 subunits); and three chaperonin proteins (DnaK, GroEL and PrsA).

Table 8 Putative highly expressed genes (see text for details) of length ≥ 200 codons in *B. subtilis* genome

Position in the genome[a]	Gene	Bias[b] All	Bias[c] RP	S3%	Function/Pathway/Subcellular location[d]
19060 +	yaaD	469	272	32.4	h.p.
119107 +	rplA	715	286	30.7	ribosomal protein L1
130683 +	fus	601	198	31.5	elongation factor G
132881 +	tufA	789	292	30.6	elongation factor Tu
135710 +	rplC	648	292	30.8	ribosomal protein L3
136367 +	rplD	690	259	29.1	ribosomal protein L4
137309 +	rplB	689	266	29.4	ribosomal protein L2
138840 +	rpsC	596	257	31.8	ribosomal protein S3
649950 +	groEL	612	303	38.1	heat–shock protein
976578 +	yhbJ	515	368	27.3	h.p.
1070718 -	prsA	431	413	39.2	molecular chaperonin
1298543 +	yjlD	568	286	35.5	h.p.
1442338 +	ykvO	519	368	27.1	h.p.
1466813 -	ykwD	453	324	33.2	h.p.
1488378 +	ykuQ	470	352	39.6	h.p.
1527731 +	pdhA	427	266	34.3	pyruvate dehydrogenase E1 alpha subunit
1528850 +	pdhB	519	237	32.7	pyruvate dehydrogenase E1 beta subunit
1529942 +	pdhC	486	221	31.7	pyruvate dehydrogenase (dihydrolipoamide acetyltransferase E2 subunit)
1531275 +	pdhD	512	222	33.1	dihydrolipoamide dehydrogenase E3 subunit
1717325 +	rpsB	624	287	33.9	ribosomal protein S2
1718167 +	tsf	628	311	31.9	elongation factor Ts
1877669 +	glnA	457	343	41.1	glutamine synthetase
2096231 -	yocJ	584	364	29.5	h.p.
2127057 +	yodC	436	439	36.8	h.p.
2235510 +	yonB	452	351	27.3	h.p.
2239580 +	yomU	431	387	26.0	h.p.
2585317 -	sodA	483	291	41.3	superoxide dismutase
2627213 -	dnaK	519	301	34.9	heat–shock protein
2886690 -	tig	537	251	30.0	trigger factor (prolyl isomerase)
2893809 -	ilvC	432	448	37.5	ketol–acid reductoisomerase, valine/isoleucine biosynthesis=
3356049 -	yurU	448	291	45.7	h.p.
3359839 -	yurY	520	416	46.5	h.p.
3361650 -	yusA	445	282	38.0	h.p.
3445139 -	yvaB	434	393	38.1	h.p.
3476910 -	eno	673	306	40.8	enolase, glycolysis
3479229 -	tpi	520	368	42.5	triose phosphate isomerase, glycolysis
3480444 -	pgk	498	403	40.5	phosphoglycerate kinase, glycolysis
3481768 -	gap	725	354	34.1	g3pd, glycolysis
3535072 +	sacB	432	366	41.5	levansucrase
3634961 -	hag	568	252	29.7	flagellin protein
3781967 -	atpD	487	391	40.5	ATP synthase (subunit beta)
3784441 -	atpA	443	403	44.3	ATP synthase (subunit alpha)
3801221 -	ywkA	457	239	33.0	h.p.
3808422 -	fbaA	573	236	32.4	fructose–1,6–bisphosphate aldolase, glycolysis
3988024 +	yxkC	484	404	24.1	h.p.

[a]Position of the translation initiation site and gene orientation (+ or −);[b]Codon bias multiplied by 1000 relative to the average *B. subtilis* gene; [c]Codon bias multiplied by 1000 relative to *B. subtilis* ribosomal proteins; [d]Abbreviations used in the table: h.p. = hypothetical protein, g3pd = glyceraldehyde-3-phosphate dehydrogenase.

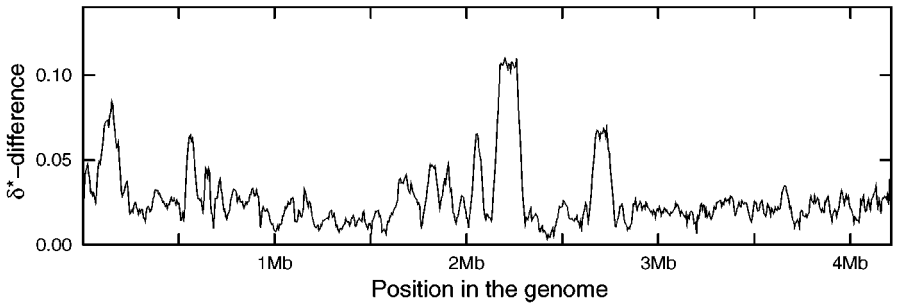


Figure 4 Plot of sliding window δ^*-differences of each 50 kb segment compared to *B. subtilis* genome signature.

This collection of highly expressed proteins parallels the highly expressed proteins of *E. coli* and *H. influenzae* (data not shown).

Sliding Window Genomic-Signature Analysis

It is useful to plot at each position for a 50-kb window a δ^*-difference compared to the average genomic signature (Figure 4). In *B. subtilis* these δ^*-difference values peak about position 2.18 Mb to 2.28 Mb. This region conatains many ORFs including many alien genes and is also the most deviant 50-kb window in amino acid usuages and in gene codon bias. The second peak extending from position 2.65 Mb to position 2.75 Mb is also abundant with ORFs and alien genes.

Pathogenicity islands (Pa-i) contain genes that cause diseases such as genes encoding invasins, adhesins, and secretion factors that often are sources of toxins. Pathogenicity islands are a subset of specialization islands (linked blocks of genes with related functions present in some closely related strains or species but not in others such as COB operon of *S. typhimurium*). These islands generally deviate sharply in G+C content from the average global genome G+C frequency. Other means of discriminating islands exploit the genomic signature profile and codon bias of the island genes compared to the genomic signature profile and codon bias relative to the average gene, respectively. We illustrate these ideas with respect to the *H. pylori* genome. The *H. pylori* genome sequence has a known pathogenicity island about 37 kb in length (*cagA*-region) (23, 29). The *cagA*-region is the most deviant in terms of genomic signature from the rest of the genome (see Figure 5). Explicitly, the average δ^*-difference between *cagA* and all other *H. pylori* genomic segments of the same length is 0.123, significantly higher than δ^*-differences among all other segments (average 0.031; range 0.006 to 0.110). In comparing the codon bias of the genes in each 50-kb segment to the average *H. pylori* gene, Figure 5 shows that the *cagA* region carries the highest codon bias.

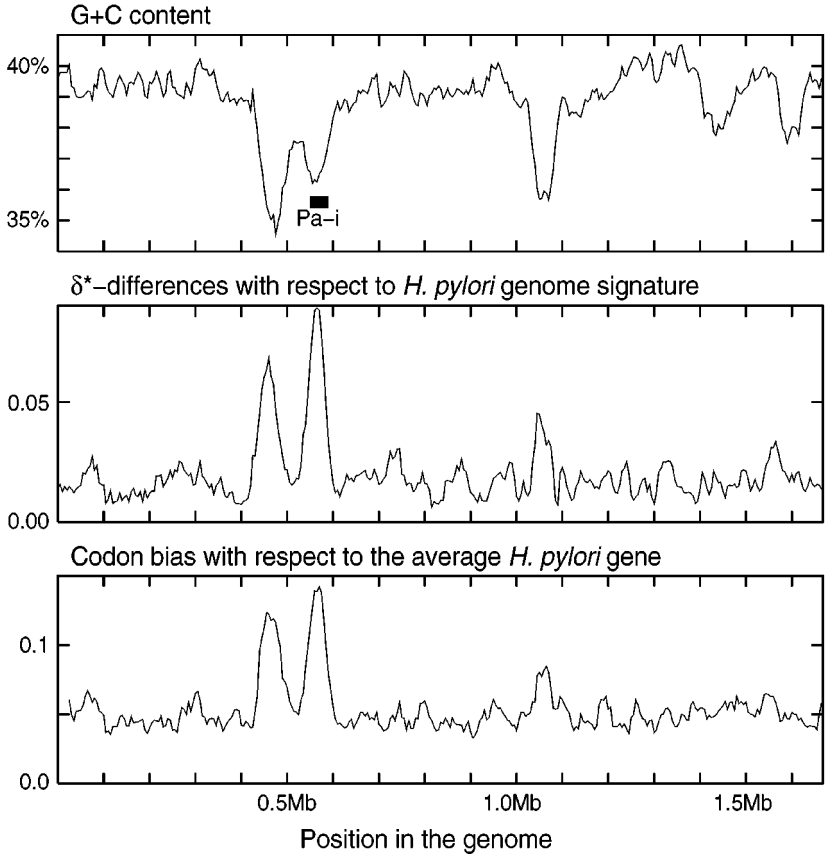


Figure 5 Local genomic characteristics in sliding windows of 50 kb in the *Helicobacter pylori* genome. Position of known pathogenicity island (Cag-region) is indicated in the top panel.

DNA DUPLEX AND COMPOSITIONAL ASYMMETRY

Several recent studies have uncovered strand compositional asymmetry between the two DNA strands in certain bacterial genomes (68, 72) (see Figure 6). A prevalence of G over C in the leading strand relative to the lagging strand was observed in the genomes of *E. coli*, of *B. subtilis*, of *M. genitalium*, and marginally of *H. influenzae*, *M. pneumoniae*, and *H. pylori*. The linear genome of *B. burgdorferi* divides into two halves of opposite G-C predominance. By contrast, dinucleotide relative abundances are approximately congruent with respect to the leading and lagging strands for all prokaryotic and eukaryotic genomes. The bias of the leading strand favoring G over C in the *E. coli*

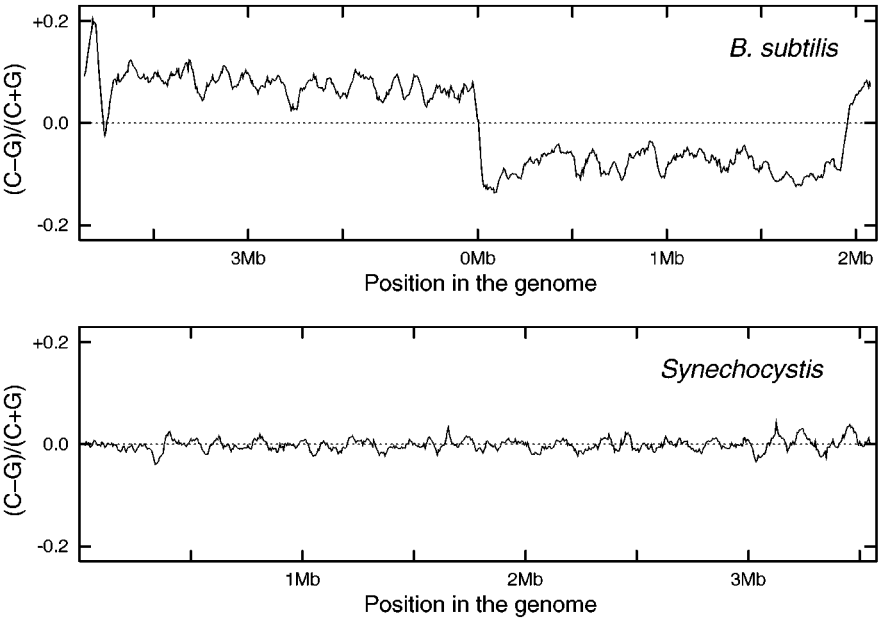


Figure 6 Sliding window plots of (C−G)/(C+G) counts for *B. subtilis* and *Synechocystis* sp. complete genomic DNA. Origin of replication in *B. subtilis* is located at 0 Mb.

genome is at variance with the common belief (e.g. 28) that large contigs of each strand in *E. coli* and most genomes tend to be approximately equal in G and C and approximately equal in A and T base content.

Strand compositional asymmetry is not observed in the cyanobacterium *Synechocystis* sp. genome nor in the archaeal genomes of *M. jannaschii*, *M. thermoautotrophicum*, *A. fulgidus*, and *P. aerophilum*. Several eukaryotic chromosomes (and long stretches) including the entire yeast (*S. cerevisiae*) genome (16 chromosomes), three chromosomes of *C. elegans*, the bithorax region (340 kb) of *D. melanogaster*, the human T-cell receptor beta locus (670 kb on chromosome 7), and the BRCA2 gene region (780 kb on chromosome 14), show no distinctive strand asymmetry.

The most consistent explanation of the data is that mononucleotide strand asymmetry in a prokaryotic genome is a consequence of a unique origin of replication coupled to bidirectional replication that favors purines (especially G > C) on the leading strand. Along these lines, strand compositional asymmetry is not apparent in the genomes of organisms known to possess multiple origins of bidirectional replication present on average about every 50 kb apart.

À propos, no origin of replication has been identified in the archaea at hand, and it has been conjectured that many archaeal genomes possess multiple origins of replication (74).

Lobry (68) associates the basis of strand compositional asymmetry to replication mutational and repair biases different in the leading versus lagging strands. Francino & Ochman (30) emphasize a mutational bias associated with transcription-coupled repair mechanisms and deamination events ($C \rightarrow T$ mutations in coding sequences). Other sources of compositional strand asymmetry might include enzymological and architectural asymmetry at the replication fork, differences in signal or binding sites in the two strands, differences in gene density coupled with amino acid and codon biases between the two strands, and dNTP pool fluctuations during the cell cycle. It appears likely that there is no single cause of the strand compositional asymmetry but rather a melange of many influences. In this context, multiple replication origins putatively attenuate strand compositional asymmetry (72).

ACKNOWLEDGMENTS

Work described in this review was supported in part by NIH grants 5R01GM10452-34, 5R01HG00335-10, and NSF grant DMS9704552 to SK and NIH grant 5R01GM51117-29 to AMC.

Literature Cited

1. Andersson SGE, Kurland CG. 1990. Codon preferences in free-living microorganisms. *Microbiol. Rev.* 54:198–210
2. Baldauf SL, Palmer JD, Doolittle WF. 1996. The root of the universal tree and the origin of eukaryotes based on elongation factor phylogeny. *Proc. Natl. Acad. Sci. USA* 93:7749–54
3. Benachenhou-Lahfa N, Forterre P, Labedan B. 1993. Evolution of glutamate dehydrogenase genes: evidence for two paralogous protein families and unusual branching patterns of the archaebacteria in the universal tree of life. *J. Mol. Evol.* 36:335–46
4. Berg DE, Howe MM. 1989. *Mobile DNA*. Washington, DC: Am Soc. Microbiol.
5. Bernardi G, Mouchiroud D, Gautier C, Bernardi G. Compositional patterns in vertebrate genomes: conservation and change in evolution. *J. Mol. Evol.* 28:7–18
6. Beutler E, Gelbart T, Han J, Koziol JA, Beutler B. 1989. Evolution of the genome and the genetic code: Selection at the dinucleotide level by methylation and polyribonucleotide cleavage. *Proc. Natl. Acad. Sci. USA* 86:192–96
7. Bhagwat AS, McClelland M. 1992. DNA mismatch correction by very short patch repair may have altered the abundance of oligonucleotides in the *Escherichia coli* genome. *Nucleic Acids Res.* 20:1663–68
8. Bird AP. 1986. CPPG-rich islands and the function of DNA methylation. *Nature* 321:209–13
9. Blackburn EH. 1991. Structure and function of telomeres. *Nature* 350:569–73
10. Blaisdell BE, Campbell AM, Karlin S. 1996. Similarities and dissimilarities of phage genomes. *Proc. Natl. Acad. Sci. USA* 93:5854–59
11. Blaisdell BE, Rudd KE, Matin A, Karlin S. 1993. Significant dispersed recurrent DNA sequences in the *Escherichia coli* genome:

several new groups. *J. Mol. Biol.* 229:833–48
12. Blumenthal AB, Kriegstein HJ, Hogness DS. 1974. The units of DNA replication in *Drosophila melanogaster* chromosomes. *Cold Spring Harbor Symp. Quant. Biol.* 38:205–23
13. Brendel V. 1996. Statistical analysis of protein sequences. In *Advances in Computational Biology*, ed. H Villar, 2:121–60. Greenwich, CT: JAI Press
14. Brendel V, Brocchieri L, Sandler SJ, Clark AJ, Karlin S. 1997. Evolutionary comparisons of RecA-like proteins across all major kingdoms of living organisms. *J. Mol. Evol.* 44:528–41
15. Brendel V, Bucher P, Nourbakhsh IR, Blaisdell BE, Karlin S. 1992. Methods and algorithms for statistical analysis of protein sequences. *Proc. Natl. Acad. Sci. USA* 89:2002–6
16. Breslauer KJ, Frank R, Blocker H, Marky LA. 1996. Predicting DNA duplex stability from the base sequence. *Proc. Natl. Acad. Sci. USA* 83:3746–50
17. Brown JR, Doolittle WF. 1997. *Archaea* and the prokaryote-to-eukaryote transtion. *Microbiol. Mol. Biol. Rev.* 61:456–502
18. Brown JR, Masuchi Y, Robb FT, Doolittle WF. 1994. Evolutionary relationships of bacterial and archaeal glutamine synthetase genes. *J. Mol. Evol.* 38:566–76
19. Bult CJ, White O, Olsen GJ, Zhou L, Fleischmann RD, et al. 1996. Complete genome sequence of the methanogenic archaeon, *Methanococcus jannaschii*. *Science* 273:1058–73
20. Burge C, Campbell AM, Karlin S. 1992. Over- and under-representation of short oligonucleotides in DNA sequences. *Proc. Natl. Acad. Sci. USA* 89:1358–62
21. Calladine CR, Drew HR. 1992. *Understanding DNA*. San Diego: Academic
22. Cardon LR, Burge C, Clayton DA, Karlin, S. 1994. Pervasive CpG suppression in animal mitochondrial genomes. *Proc. Natl. Acad. Sci. USA* 91:3799–803
23. Covacci A, Falkow S, Berg DE, Rappuoli R. 1997. Did the inheritance of a pathogenicity island modify the virulence of *Helicobacter pylori*? *Trends Microbiol.* 5:205–8
24. Delcourt SG, Blake RD. 1991. Stacking energies in DNA. *J. Biol. Chem.* 266:15160–69
25. Doerfler W. 1983. DNA methylation and gene activity. *Annu. Rev. Biochem.* 52:93–124
26. Echols H, Goodman MF. 1991. Fidelity mechanisms in DNA replication. *Annu. Rev. Biochem.* 60:477–511
27. Fickett JW. 1982. Recognition of protein coding regions in DNA sequences. *Nucleic Acids Res.* 10:5303–18
28. Fickett JW, Torney DC, Wolf DR. 1992. Base compositional structure of genomes. *Genomics* 13:1056–64
29. Finley BB, Falkow S. 1997. Common themes in microbial pathogenicity II. *Mol. Biol. Microbiol. Rev.* 61:136–69
30. Francino MP, Ochman H. 1997. Strand asymmetries in DNA evolution. *Trends Genet.* 13:240–45
31. Gillespie JH. 1991. *The Causes of Molecular Evolution*. New York: Oxford Univ. Press
32. Gilson E, Saurin W, Perrin D, Bachellier S, Hofnung M. 1991. Palindromic units are part of a new bacterial interspersed mosaic element (BIME). *Nucleic Acids Res.* 19:1375–83
33. Gupta RS, Golding GB. 1993. Evolution of HSP70 gene and its implications regarding relationships between archaebacteria, eubacteria, and eukaryotes. *J. Mol. Evol.* 37:573–82
34. Gupta RS, Golding GB. 1996. The origin of the eukaryotic cell. *Trends Biochem. Sci.* 21:166–71
35. Gupta RS, Singh B. 1994. Phylogenetic analysis of 70 kD heat shock protein sequences suggests a chimeric origin for the eukaryotic cell nucleus. *Curr. Biol.* 4:1104–14
36. Hanawalt PC. 1994. Transcription-coupled repair and human disease. *Science* 266: 1957–58
37. Hess ST, Blake JD, Blake RD. 1994. Wide variations in neighbor-dependent substitution rates. *J. Mol. Biol.* 236:1022–33
38. Holmquist GP, Filipski J. 1994. Organization of mutations along the genome: a prime determinant of genome evolution. *Trends Ecol. Evol.* 9:65–69
39. Hunter CA. 1993. Sequence-dependent DNA-structure: the role of base stacking interactions. *J. Mol. Biol.* 230:1025–54
40. Inman RB. 1966. A denaturation map of the lambda phage DNA moleculae determined by electron microscopy. *J. Mol. Biol.* 18:464–76
41. Josse J, Kaiser AD, Kornberg A. 1961. Enzymatic synthesis of deoxyribonucleic acid. VIII. Frequencies of nearest neighbor base sequences in deoxyribonucleic acid. *J. Biol. Chem.* 263:864–75
42. Karlin S. 1997. *Assessing Inhomogeneities in Bacterial Long Genomic Sequences*. Santa Fe, NM: RECOMB 97
43. Karlin S, Brendel V. 1992. Chance and statistical significance in protein and DNA sequence analysis. *Science* 257:39–49

44. Karlin S, Burge C. 1995. Dinucleotide relative abundance extremes: a genomic signature. *Trends Genet.* 11:283–90
45. Karlin S, Burge C, Campbell AM. 1992. Statistical analyses of counts and distributions of restriction sites in DNA sequences. *Nucleic Acids Res.* 20:1363–70
46. Karlin S, Campbell AM. 1994. Which bacterium is the ancestor of the animal mitochondrial genome? *Proc. Natl. Acad. Sci. USA* 91:12842–46
47. Karlin S, Cardon L. 1994. Computational DNA sequence analysis. *Annu. Rev. Microbiol.* 48:619–54
48. Karlin S, Doerfler W, Cardon LR. 1994. Why is CpG suppressed in the genomes of virtually all small eukaryotic viruses but not in those of large eukaryotic viruses? *J. Virol.* 68:2889–97
49. Karlin S, Ladunga I. 1994. Comparisons of eukaryotic genomic sequences. *Proc. Natl. Acad. Sci. USA* 91:12832–36
50. Karlin S, Leung M-Y. 1991. Some limit theorems on distributional patterns of balls in urns. *Ann. Appl. Probl.* 4:152–67
51. Karlin S, Mocarski E, Schachtel GA. 1994. Molecular evolution of herpesviruses: genomic and protein sequence comparisons. *J. Virol.* 68:1886–902
52. Karlin S, Mrázek J. 1996. What drives codon choices in human genes? *J. Mol. Biol.* 262:459–72
53. Karlin S, Mrázek J. 1997. Prokaryotic genome-wide comparisons and evolutionary implications. In *Bacterial Genomes: Physical Structure and Analysis*, ed. FJ de Bruijn, GM Weinstock, JR Lupski, pp. 196–212. New York: Chapman & Hall
54. Karlin S, Mrázek J. 1997. Compositional differences within and between eukaryotic genomes. *Proc. Natl. Acad. Sci. USA* 94:10227–32
55. Karlin S, Mrázek J, Campbell AM. 1996. Frequent oligonucleotides and peptides of the *Haemophilus influenzae* genome. *Nucleic Acids Res.* 24:4263–72
56. Karlin S, Mrázek J, Campbell AM. 1997. Compositional biases of bacterial genomes and evolutionary implications. *J. Bacteriol.* 179:3899–913
57. Karlin S, Mrázek J, Campbell AM. 1998. Codon usages in different gene classes of the *E. coli* genome. *Mol. Microbiol.* In press
58. Kimura M. 1983. *The Neutral Theory of Molecular Evolution*. New York: Cambridge Univ. Press
59. Klenk HP, Clayton RA, Tomb JF, White O, Nelson KE, et al. 1997. The complete genome sequence of the hyperthermophilic, sulfate-reducing archaeon *Archaeoglobus fulgidus*. *Nature* 390:364–70
60. Koonin EV, Mushegian AR, Rudd KE. 1996. Sequencing and analysis of bacterial genomes. *Curr. Biol.* 6:404–16
61. Krawiec S, Riley M. 1990. Organization of the bacterial chromosome. *Microbiol. Rev.* 54:502–39
62. Krysan PJ, Smith JG, Calos MP. 1993. Autonomous replication in human cells of multimers of specific human and bacteria DNA sequences. *Mol. Cell. Biol.* 13:2688–96
63. Kunkel TA. 1992. Biological asymmetries and the fidelity of eukaryotic DNA replication. *BioEssays* 14:303–8
64. Kurland CG. 1993. Major codon preference: theme and variations. *Biochem. Soc. Trans.* 21:841–46
65. Lake JA. 1989. Origin of the eukaryotic nucleus: eukaryotes and eocytes are genotypically related. *Can. J. Microbiol.* 35:109–18
66. Lawrence JG, Ochman H. 1997. Amelioration of bacterial genomes: rates of change and exchange. *J. Mol. Evol.* 44:383–97
67. Lawrence JG, Roth JR. 1996. Selfish operons: horizontal transfer may drive the evolution of gene clusters. *Genetics* 143:1843–60
68. Lobry JR. 1996. Asymmetric substitution paterns in the two DNA strands of bacteria. *Mol. Biol. Evol.* 13:660–65
69. Lorenz MG, Wackernagel W. 1994. Bacterial gene transfer by natural genetic transformation in the environment. *Microbiol. Rev.* 58:563–602
70. Médigue C, Rouxel T, Vigier P, Henaut A, Danchin A. 1991. Evidence for horizontal gene transfer in *Escherichia coli* speciation. *J. Mol. Biol.* 222:851–56
71. Mrázek J, Karlin S. 1996. A new significant recurrent dyad pairing in *Haemophilus influenzae*. *Trends Biochem. Sci.* 21:201–2
72. Mrázek J, Karlin S. 1998. Strand compositional asymmetry in bacterial and large viral genomes. *Proc. Natl. Acad. Sci. USA* 95:3720–25
73. Nelson M, McClelland M. 1991. Site-specific methylation: effect on DNA modification methyltransferases and restriction endonucleases. *Nucleic Acids Res.* 19:2045–75
74. Olsen GJ, Woese CR. 1997. Archaeal genomes: an overview. *Cell* 89:991–94
75. Otwinowski Z, Schevitz RW, Zhang RG, Lawson CL, Joachimiak A, et al. 1988. Crystal structure of Trp repressor operator complex at atomic resolution. *Nature* 335:321–29
76. Rafferty JB, Somers WS, StGirons I, Phillips SEV. 1989. 3-dimensional crystal-

structures of *Escherichia coli* Met repressor with and without corepressor. *Nature* 341:705–10
77. Riley M. 1993. Functions of the gene products of *Escherichia coli*. *Microbiol. Rev.* 57:862–952
78. Rivera MC, Lake JA. 1992. Evidence that eukaryotes and eocyte prokaryotes are immediate relatives. *Science* 257:74–76
79. Robinson NJ, Robinson PJ, Gupta A, Bleasby AJ, Whitton BA, Morby AP. 1995. Singular over-representation of an octameric palindrome, HIP1, in DNA from many cyanobacteria. *Nucleic Acids Res.* 23:729–35
80. Russell GJ, Subak-Sharpe JH. 1977. Similarity of the general designs of protochordates and invertebrates. *Nature* 266:533–35
81. Russell GJ, Walker PM, Elton RA, Subak-Sharpe JH. 1976. Doublet frequency analysis of fractionated vertebrate nuclear DNA. *J. Mol. Biol.* 108:1–23
82. Sandler SJ, Satin LH, Samra HS, Clark AJ. 1996. RecA-like genes from three archaean species with putative protein products similar to Rad51 and Dmc1 proteins of the yeast *Saccharomyces cerevisiae*. *Nucleic Acids Res.* 24:2125–32
83. Selker EU. 1990. Premeiotic instability of repeated sequences in *Neurospora crassa*. *Annu. Rev. Genet.* 24:579–613
84. Sharp PM, Li WH. 1987. The codon adaptation index: a measure of directional synonymous codon usage bias, and its potential applications. *Nucleic Acids Res.* 15:1281–95
85. Sharp PM, Matassi G. 1994. Codon usage and genome evolution. *Curr. Opin. Genet. Dev.* 4:851–60
86. Smith DR, Doucette-Stamm LA, Deloughery C, Lee H, Dubois J, et al. 1997. Complete genome sequence of Methanobacterium thermoautotrophicum DELTA-H: functional analysis and comparative genomics. *J. Bacteriol.* 179:7135–55
87. Smith HO, Tomb J-F, Dougherty BA, Fleischmann RD, Venter JC. 1995. Frequency and distribution of DNA uptake signal sequences in the *Haemophilus influenzae* Rd genome. *Science* 269:538–40
88. Strauss EJ, Falkow S. 1997. Microbial pathogenesis: genomics and beyond. *Science* 276:707–12
89. Tatusov RL, Mushegian AR, Bork P, Brown NP, Hayes WS, et al. 1996. Metabolism and evolution of *Haemophilus influenzae* deduced from a whole-genome comparison with *Escherichia coli*. *Curr. Biol.* 6:279–91
90. Tazi J, Bird A. 1990. Alternative chromatin structure at CpG islands. *Cell* 60:909–20
91. Travers AA. 1993. *DNA-Protein Interactions*. New York: Chapman & Hall
92. Willard HF, Waye JS. 1987. Hierarchical order in chromosome-specific human alpha-satellite DNA. *Trends Genet.* 3:192–98
93. Woese CR, Kandler O, Wheelis ML. 1990. Towards a natural system of organisms: proposal for the domains Archaea, Bacteria, and Eukarya. *Proc. Natl. Acad. Sci. USA* 87:4576–79