JOURNAL OF BACTERIOLOGY, June 1997, p. 3899–3913 Vol. 179, No. 12
0021-9193/97/$04.00+0

Compositional Biases of Bacterial Genomes and Evolutionary Implications

SAMUEL KARLIN,[1]* JAN MRÁZEK,[1] AND ALLAN M. CAMPBELL[2]

Department of Mathematics, Stanford University, Stanford, California 94305-2125,[1] *and Department of Biological Sciences, Stanford University, Stanford, California 94305-5020*[2]

Received 5 December 1996/Accepted 9 April 1997

We compare and contrast genome-wide compositional biases and distributions of short oligonucleotides across 15 diverse prokaryotes that have substantial genomic sequence collections. These include seven complete genomes (*Escherichia coli*, *Haemophilus influenzae*, *Mycoplasma genitalium*, *Mycoplasma pneumoniae*, *Synechocystis* sp. strain PCC6803, *Methanococcus jannaschii*, and *Pyrobaculum aerophilum*). A key observation concerns the constancy of the dinucleotide relative abundance profiles over multiple 50-kb disjoint contigs within the same genome. (The profile is $\rho^*_{XY} = f^*_{XY}/f^*_X f^*_Y$ for all *XY*, where f^*_X denotes the frequency of the nucleotide *X* and f^*_{XY} denotes the frequency of the dinucleotide *XY*, both computed from the sequence concatenated with its inverted complementary sequence.) On the basis of this constancy, we refer to the collection $\{\rho^*_{XY}\}$ as the genome signature. We establish that the differences between $\{\rho^*_{XY}\}$ vectors of 50-kb sample contigs of different genomes virtually always exceed the differences between those of the same genomes. Various di- and tetranucleotide biases are identified. In particular, we find that the dinucleotide CpG=CG is underrepresented in many thermophiles (e.g., *M. jannaschii*, *Sulfolobus* sp., and *M. thermoautotrophicum*) but overrepresented in halobacteria. TA is broadly underrepresented in prokaryotes and eukaryotes, but normal counts appear in *Sulfolobus* and *P. aerophilum* sequences. More than for any other bacterial genome, palindromic tetranucleotides are underrepresented in *H. influenzae*. The *M. jannaschii* sequence is unprecedented in its extreme underrepresentation of CTAG tetranucleotides and in the anomalous distribution of CTAG sites around the genome. Comparative analysis of numbers of long tetranucleotide microsatellites distinguishes *H. influenzae*. Dinucleotide relative abundance differences between bacterial sequences are compared. For example, in these assessments of differences, the cyanobacteria *Synechocystis*, *Synechococcus*, and *Anabaena* do not form a coherent group and are as far from each other as general gram-negative sequences are from general gram-positive sequences. The difference of *M. jannaschii* from low-G+C gram-positive proteobacteria is one-half of the difference from gram-negative proteobacteria. Interpretations and hypotheses center on the role of the genome signature in highlighting similarities and dissimilarities across different classes of prokaryotic species, possible mechanisms underlying the genome signature, the form and level of genome compositional flux, the use of the genome signature as a chronometer of molecular phylogeny, and implications with respect to the three putative eubacterial, archaeal, and eukaryote domains of life and to the origin and early evolution of eukaryotes.

In this report, we describe measures of genomic similarities that do not depend on prior alignment of homologous sequences and apply them to sufficiently large samples of prokaryotic genomic sequences. The approach departs from almost all other methods of similarity analysis and evolutionary reconstruction by using as its basis sequence information derived from the entire genome rather than individual genes. Comparisons are based on DNA sequence relative abundance values of di- and tetranucleotides. These measurements appear to discriminate local DNA conformational tendencies that are constant throughout the genome. Factors that can influence DNA structure include dinucleotide stacking stability, constraints on helicity, methylation modifications, context-dependent mutation pressures, and DNA replication and repair mechanisms (see Discussion). Genomic sequences are analyzed with respect to similarities and differences of relative abundance values of short oligonucleotides within and between genomes. In particular, our analysis centers on comparisons and contrasts of compositional extremes and short oligonucleotide distributional anomalies across 15 substantial prokaryotic sequence aggregates. The primary data include genomic collections from sequences of five gram-negative proteobacteria (including two complete genomes), three gram-positive bacteria, two mycoplasmas (both complete genomes), two cyanobacteria (one complete genome), and three thermophilic archaea (one complete genome) (Table 1).

Genomic sequences display internal heterogeneity of many kinds, including G+C variation, isochore compartments, coding versus noncoding, mobile insertion sequences, methylation patterns, recombinational hot spots, and hierarchies of repeats. Collectively, the dinucleotide relative abundance values $\{\rho^*_{XY}$; see Materials and Methods$\}$ calculated, for example, for disjoint 50-kb contigs covering the genome, give each genome a signature that is approximately constant throughout the genome (6, 32, 35). Along these lines, our recent studies have demonstrated that the dinucleotide relative abundance values of different sequence samples of DNA from the same organism are generally much more similar to each other than they are to corresponding sequence samples from other organisms and that closely related organisms generally have more similar dinucleotide relative abundance values than do distantly related organisms (32, 36). These highly stable normalized DNA-

* Corresponding author.

TABLE 1. DNA sequence data sets

Sequence	Length	% of complete genome	Note	G+C content (%)	Type
Escherichia coli	4.639 Mb	100		51	γ-Proteobacteria
Haemophilus influenzae	1.830 Mb	100		38	γ-Proteobacteria
Mycoplasma genitalium	580 kb	100		32	Putatively derived from low-G+C gram-positive sequence
Mycoplasma pneumoniae	816 kb	100		40	
Synechocystis sp.	3.573 Mb	100		48	Cyanobacteria
Methanococcus jannaschii	1.665 Mb	100		31	Archaea (thermophile)
Pyrobaculum aerophilum	2.172 Mb	99	11 contigs	51	Archaea (thermophile)
Bacillus subtilis	508 kb	13	3 contigs including 276 kb about replication origin	44	Gram positive
Salmonella typhimurium	407 kb		94 GenBank entries ≥ 2.5 kb	52	γ-Proteobacteria
Pseudomonas aeruginosa	298 kb		78 GenBank entries ≥ 2.5 kb	64	γ-Proteobacteria
Rhizobium meliloti	246 kb		51 GenBank entries ≥ 2.5 kb	61	α-Proteobacteria
Staphylococcus aureus	298 kb		70 GenBank entries ≥ 2.5 kb	34	Gram positive
Lactococcus lactis	242 kb		52 GenBank entries ≥ 2.5 kb	35	Gram positive
Synechococcus sp.	187 kb		47 GenBank entries ≥ 2.5 kb	53	Cyanobacteria
Methanobacterium thermoautotrophicum	175 kb		48 nonredundant GenBank entries ≥ 1 kb	48	Archaea (thermophile)

doublet frequencies suggest that there may be genome-wide factors such as functions of the replication and repair machinery, context-dependent mutation patterns, and conformational tendencies of double-stranded dinucleotides (base steps) that impose limits on the compositional and structural variations of any particular genomic sequence and that the set of dinucleotide relative abundance values reflects the influence of these organism-dependent factors.

Dinucleotide relative abundance profiles $\{\rho^*_{XY}$ for all $XY\}$ (designated the genome signatures) are equivalent to the general designs derived from biochemical nearest-neighbor frequency analysis that were evaluated extensively three decades ago in samples of genomic DNA from many organisms (29, 57, 58). It was observed that the set of dinucleotide relative abundance values is essentially constant throughout a genome: for bulk genomic DNA, for DNA fractions differing in sequence complexity (renaturation rate fractions), for euchromatin or heterochromatin, and for distinct base compositional (density gradient) fractions of nuclear DNA (58). In reference 38, we introduced the codon signature, defined as the dinucleotide relative abundances at the distinct codon positions {1, 2}, {2, 3}, and {3, 4} (4 = 1 of the next codon). For large collections of genes (50 or more), we found that the codon signature, like the genome signature, is essentially an invariant. Moreover, the codon signature largely adheres to the genome signature but accommodates amino acid constraints (38).

In this report, we firmly corroborate robustness of the genome signature of the 15 bacterial sequences. Whereas the signature as defined includes the relative abundances of all dinucleotides, it is most markedly influenced by those particular dinucleotides that are either extremely overrepresented or extremely underrepresented in a given genome. We therefore identify certain dinucleotides that are extremely over- or underrepresented across the 15 genomes, either broadly or in particular species, and include some discussion of how such extremes may come about. We analyze tetranucleotides in the same manner. We also address the following questions. Are there strong compositional influences due to factors such as restriction systems, methylation modifications, insertion sequences, and membrane attachment sites? What are possible mechanisms underlying the genome signature? How may similarities and differences of di- and tetranucleotide compositional extremes among classes of organisms provide insights into evolutionary relationships?

MATERIALS AND METHODS

Data. Data encompass the seven complete genomes, i.e., *Escherichia coli* (4.6 Mb), *Haemophilus influenzae* (1.83 Mb), *Mycoplasma genitalium* (580 kb), *Mycoplasma pneumoniae* (816 kb), *Methanococcus jannaschii* (1.67 Mb), *Synechocystis* sp. strain PCC6803 (3.57 Mb), and *Pyrobaculum aerophilum* (2.2 Mb), and large sequence collections of *Bacillus subtilis* (508 kb; three contigs including one of 276 kb centered at *ori* C) and *Salmonella typhimurium* (407 kb). In several discussions, we augment the data to include other prokaryote nonredundant genomic collections of aggregate exceeding about 200 kb where each individual contributing sequence constitutes a contiguous piece of at least 2.5 kb (Table 1).

Relative abundance extremes. Dinucleotide contrasts are commonly assessed through the odds ratio functional $\rho_{XY} = f_{XY}/f_X f_Y$, where f_X denotes the frequency of the nucleotide X and f_{XY} is the frequency of the dinucleotide XY in the sequence under study. For double-stranded DNA sequences, a symmetrized version ρ^*_{XY} is computed from frequencies of the sequence concatenated with its inverted complementary sequence (11). Dinucleotide relative abundances ρ^*_{XY} effectively assess contrasts between the observed dinucleotide frequencies and those that are expected from the component mononucleotide frequencies. Statistical theory and data experience indicate conservative estimates, $\rho^*_{XY} \geq 1.23$ or ≤ 0.78, when the doublet XY is of significantly high or low relative abundance, respectively (see the footnote to Table 2 for more refined criteria of discrimination). We refer to values in the range $0.78 < \rho^*_{XY} < 1.23$ as normal.

The corresponding third- and fourth-order oligonucleotide measures which factor out all lower-order biases are

$$\gamma^*_{XYZ} = (f^*_{XYZ} f^*_X f^*_Y f^*_Z)/(f^*_{XY} f^*_{YZ} f^*_{XNZ})$$

and

$$\tau^*_{XYZW} = (f^*_{XYZW} f^*_{XY} f^*_{XNZ} f^*_{XN_1N_2W} f^*_{YZ} f^*_{YNW} f^*_{ZW})/(f^*_{XYZ} f^*_{XYNW} f^*_{XNZW} f^*_{YZW} f^*_X f^*_Y f^*_Z f^*_W),$$

respectively, where N is any nucleotide and W, X, Y, Z are each one of A, C, G, or T (35). These above formulas assess log regression contingency interactions (5). Markov calculations of biases are determined by the formulas (53)

$$\tilde{\gamma}_{XYZ} = \frac{f^*_{XYZ} f^*_Y}{f^*_{XY} f^*_{YZ}} \text{ and } \tilde{\tau}_{XYZW} = \frac{f^*_{XYZW} f^*_{YZ}}{f^*_{XYZ} f^*_{YZW}}.$$

Dinucleotide relative abundance differences. We summarize the difference (dissimilarity) between two sequences f and g (from different organisms or from different regions of a single genome) by the average absolute difference of the dinucleotide relative abundance values

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

where the sum extends over all dinucleotides.

TABLE 2. Dinucleotide relative abundance extremes (genome signatures), dinucleotide relative abundance ranges for multiple ~50-kb sequence samples, and G+C content

Sequence	Dinucleotide relative abundance extremes[a]								G+C content (%)
	CG	GC	TA	AT	CC GG	TT AA	AC GT	GA TC	
E. coli		+	−						51
Overall ρ*		1.28	0.75						
Range (93 samples)		1.20–1.35	0.69–0.81						
H. influenzae		++	−			+			38
Overall ρ*		1.43	0.75			1.25			
Range (36 samples)		1.28–1.57	0.70–0.80			1.20–1.29			
M. genitalium	−−−		−	−					32
Overall ρ*	0.39		0.75	0.77					
Range (12 samples)	0.32–0.50		0.71–0.78	0.72–0.81					
M. pneumoniae			−	−		+			40
Overall ρ*			0.77	0.71		1.30			
Range (16 samples)			0.74–0.81	0.65–0.74		1.26–1.35			
Synechocystis	−		−		++	++			48
Overall ρ*	0.75		0.75		1.36	1.32			
Range (71 samples)	0.71–0.81		0.71–0.79		1.20–1.40	1.24–1.37			
M. jannaschii	−−−				++		−		31
Overall ρ*	0.32				1.38		0.72		
Range (33 samples)	0.24–0.50				1.31–1.45		0.68–0.77		
P. aerophilum									51
B. subtilis		+	−−			+	−		44
Overall ρ*		1.24	0.66			1.23	0.76		
Range (10 samples)		1.14–1.33	0.61–0.71			1.20–1.27	0.69–0.81		
S. typhimurium		+							52
Overall ρ*		1.29							
Range (8 samples)		1.22–1.34							
P. aeruginosa			−−						64
Overall ρ*			0.56						
Range (6 samples)			0.52–0.61						
R. meliloti	+		−−	++				+	61
Overall ρ*	1.27		0.50	1.33				1.24	
Range (5 samples)	1.26–1.29		0.47–0.52	1.27–1.39				1.23–1.25	
S. aureus									34
L. lactis			−						35
Overall ρ*			0.72						
Range (5 samples)			0.69–0.77						
Synechococcus			−−						53
Overall ρ*			0.59						
Range (4 samples)			0.56–0.62						
M. thermoautotrophicum	−−		−		+				48
Overall ρ*	0.53		0.73		1.23				
Range (4 samples)	0.50–0.55		0.70–0.75		1.19–1.25				

[a] Overrepresentation is indicated by + ($1.23 \leq \rho^* < 1.30$), ++ ($1.30 \leq \rho^* < 1.50$), and +++ ($1.50 \leq \rho^*$); underrepresentation is indicated by − ($0.70 < \rho^* \leq 0.78$), −− ($0.50 < \rho^* \leq 0.70$), and −−− ($\rho^* \leq 0.50$). + is a statistically significant extreme which would occur in a random 50-kb double-stranded sequence with probability $P^* \leq 10^{-3}$, for ++ $P^* \leq 10^{-6}$, and for +++ $P^* \leq 10^{-9}$. The dinucleotide relative abundances of $\frac{TG}{CA}$ and $\frac{AG}{CT}$ show no extremes.

Partial orderings. To avoid being misled by a few extreme dinucleotide relative abundances that exert a large influence on the value of δ^* (*f*,*g*), we invoke a method of partial orderings where each sequence is represented by the vector of its 16 dinucleotide relative abundances (ρ^*_{XY}). The dinucleotide relative abundance vectors of the two sequences are compared with a corresponding 16-component vector of a sequence standard $\mathcal{S}$. If one of the two sequences $\mathcal{A}$ and $\mathcal{B}$, say $\mathcal{A}$, is closer to the standard $\mathcal{S}$ in at least 14 of the 16 components, a dominance ordering between the two genomes relative to the standard is determined, expressed as $\mathcal{A}$ dominates $\mathcal{B}$ (35). These evaluations relative to the standard provide a partial ordering among the sequences. Because the partial ordering depends only on how many of the 16 dinucleotides are closer to the standard (and not on how much closer), all 16 dinucleotides are given equal weight in each comparison. For a given standard, the closest sequences are those which are undominated and dominate several other sequences; the most distant sequences are those that are dominated by several sequences but dominate none. With each standard, the comparisons are made for every pair of sequences.

Analysis of the distribution of marker arrays. A general problem arises of how to characterize significant irregularities (clustering, overdispersion, or excessive evenness) in the spacings of a marker array along a sequence of nucleotides or amino acids. Anomalies in the distribution of a marker array can be ascertained in two equivalent ways: *r*-scan statistics and sliding window counts (35). In particular, analysis of spacings ensues by consideration of the cumulative lengths of *r* successive distances between the markers, where $R_i^{(r)}$ is the distance between marker *i* and marker $i + r$, called *r*-scan lengths, and *r* is a parameter. The lengths of the shortest and longest *r* scans are appropriate statistics for detecting cases of significant clustering, significant overdispersion, or excessive regularity in the spacings of the marker. In this context, we compare the theoretical distribution of $m_r^* = \min_i R_i^{(r)}$ and $M_r^* = \max_i R_i^{(r)}$ calculated under a random model with the observed *r*-scan lengths. Probabilistic formulas have been developed (17, 37). The case for $r = 1$ is classical. By varying *r*, organization on different scales can be detected. For other applications of *r*-scan statistics, see references 31 and 35.

RESULTS

Dinucleotide compositional extremes. Each sequence (genome) is partitioned into disjoint 50-kb contigs generating an array of contigs. Table 2 summarizes the dinucleotide relative abundance extremes for the bacterial sequence collections. The limited range of the extreme ρ^*_{XY} values over 50-kb samples confirms the substantial invariance around the genome of

the dinucleotide relative abundance profile. (The results are basically congruent and even more stable for larger contig size, e.g., 100 kb.) There are clear trends, as follows.

(i) The dinucleotide TpA=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* ($\rho^*_{TA} \approx 1.07$) and *Sulfolobus* sp. ($\rho^*_{TA} \approx 1.01$) (39). TA underrepresentation is also pervasive in eukaryotic species sequences, although not in eukaryotic viral genomes or in mitochondrial and chloroplast genomes (34).

(ii) GC is predominantly high in γ-proteobacterial sequences, in many β-proteobacterium examples, and in several low-G+C gram-positive bacterial genomes (e.g., *B. subtilis* and *Streptococcus mutans*).

(iii) CG is underrepresented in *M. genitalium* to the same extent as in vertebrate DNA. The same holds for *Mycoplasma capricolum* (35) but not for *Mycoplasma pneumoniae*. CG is also underrepresented in the low-G+C gram-positive sequences of *Streptococcus* and *Clostridium* (e.g., *Streptococcus pneumoniae*, *Streptococcus mutans*, and *Clostridium perfringens* [see Table 9]) and in many thermophiles, including *M. jannaschii*, *Sulfolobus* sp., *Methanobacterium thermoautotrophicum*, and *Thermus* sp. At the other extreme, CG overrepresentations stand out in *Bacillus stearothermophilus*, in halophiles, and also in several α- and β-proteobacterial genomes (e.g., *Rhizobium* sp. and *Neisseria gonorrhoeae*).

(iv) AT is overrepresented in most α-proteobacterial sequences (see Table 2 for *Rhizobium meliloti*).

(v) There are few bacterial genomic sequences devoid of any dinucleotide extremes. In this vein, *S. aureus* and the archaeon *P. aerophilum* show all dinucleotide relative abundances in the random range (Table 2). Also, the cyanobacterium *Anabaena* sp. sequences entail no dinucleotide biases (data not shown). On the other side, in the *Synechocystis* genome, four dinucleotides are over or underrepresented (Table 2).

Tetranucleotide compositional extremes. Table 3 displays over- and underrepresented tetranucleotides in the prokaryotic genomes under study. The τ* range for each tetranucleotide was determined for a partition of each genome into 100-kb contigs, and tetranucleotide extremes common to all samples are reported in Table 3. Strong tetranucleotide biases appear foremost in *M. jannaschii* and *H. influenzae*, and the strongest biases consist of avoidance of certain palindromic tetranucleotides. *M. genitalium* and *M. pneumoniae* entail few biases on the tetranucleotide level. Tetranucleotide compositions of *B. subtilis* are fully in the random range.

H. influenzae. Among the prokaryotes whose genomes are studied, *H. influenzae* is striking for breadth and types of underrepresented palindromic tetranucleotides. Thus, 9 of the 16 possible are significantly underrepresented ($\tau^* \leq 0.78$) and five other tetranucleotides are overrepresented, each of them differing by a single base substitution from at least one of the low palindromic tetranucleotides (Table 3). It is tempting to explain many of these underrepresentations by restriction avoidance. (The tetranucleotide CTAG warrants special consideration and will be discussed separately.) Many restriction enzymes are found in *Haemophilus* species, including *Hpa*II (target CCGG, $\tau^* = 0.37$), *Hfp*2 (CATG, $\tau^* = 0.43$), *Hae*III (GGCC, $\tau^* = 0.50$) {*Hha*I, *Hin*dp1} (GCGC, $\tau^* = 0.62$), and *Hin* 1056I (CGCG, $\tau^* = 0.70$). We do not know if the tetranucleotides ACGT and TCGA are restriction sites. Other underrepresented palindromic tetranucleotides are embedded in six-palindrome restriction sites, specifically *Hae*II (RGCGCY), *Hpa*I (GTTAAC), *Hin*1I (GRCGYC), and *Hin*dII (GTYRAC). However, restriction sites are not underrepresented in all bacteria. For example, in *B. subtilis*, CCGG, GGCC, and CGCG are established target sites for the restriction systems *Bsu*FI, *Bsu*RI, and *Bsu*EII, respectively, but carry normal relative abundance values ($\tau^* = 0.93$, 0.94, and 0.94, respectively).

TABLE 3. Tetranucleotide extremes (τ* values)[a]

	Low		High	
***E. coli*:**[b]				
	CTAG	0.24	none	
	CAAG/CTTG	0.73		
***H. influenzae*:**[b]				
	CCGG	0.37	CCGC/GCGG	1.25
	CATG	0.43		
	GGCC	0.50	CAGG/CCTG	1.33
	GCGC	0.62		
	CTAG	0.63	GACC/GGTC	1.26
	CGCG	0.70		
	TATA	0.71	GCCC/GGGC	1.40
	TCGA	0.78		
	GTAC	0.78	CTAA/TTAG	1.24
***M. genitalium*:**				
	CCGA/TCGG	0.76	none	
	TATA	0.78		
***M. pneumoniae*:**				
	TATA	0.78	none	
***Synechocystis sp*:**				
	CGCG	0.37	CGTA/TACG	1.27
	GCGC	0.63		
***M. jannaschii*:**[b]				
	CTAG	0.06	CCAG/CTGG	1.32
	GATC	0.11		
	GTAC	0.31	GAAC/GTTC	1.35
	GCGC	0.32		
	GGCC	0.43	GACC/GGTC	1.39
	CCGG	0.77		
	CAGG/CCTG	0.77		
***P. aerophilum*:**				
	CAGG/CCTG	0.77	GATC	1.34
***B. subtilis*:**				
	none		none	
***S. typhimurium*:**[b]				
	CTAG	0.32	none	
	CAAG/CTTG	0.74		
***P. aeruginosa*:**[b]				
	CTAG	0.35	CTAC/GTAG	1.36
			ATAG/CTAT	1.25
			TTAA	1.29
***R. meliloti*:**[b]				
	CTAG	0.45	ATAG/CTAT	1.35
***S. aureus*:**				
	GGCC	0.75	none	
***L. lactis*:**				
	GATC	0.52	none	
	GGCC	0.67		
***Synechococcus sp.*:**				
	TATA	0.75	none	
***M. thermoautotrophicum*:**[b]				
	CTAG	0.29	ATAG/CTAT	1.32
			CGAA/TTCG	1.25

[a] See Materials and Methods for definition; all significance evaluations are based on double-stranded DNA.
[b] The linked tetranucleotides differ by a single base change.

How are the foregoing rare tetranucleotide palindromes distributed in the *H. influenzae* genome? Figure 1 displays the tetranucleotide relative abundance values of CCGG sites for all sequence windows of length 30 kb with 3-kb shifts. CCGG is the most underrepresented tetranucleotide of *H. influenzae* ($\tau^*_{CCGG} = 0.37$). Highest peak occurrences (relative clusters) of CCGG are in rRNA operons and in the region of the Mu-like phage sequences of *H. influenzae*. The corresponding display for GGCC sites ($\tau^*_{GGCC} = 0.50$) again peak in rRNA genes (data not shown). No unusual distribution is seen for sliding window τ* values of the underrepresented tetranucle-

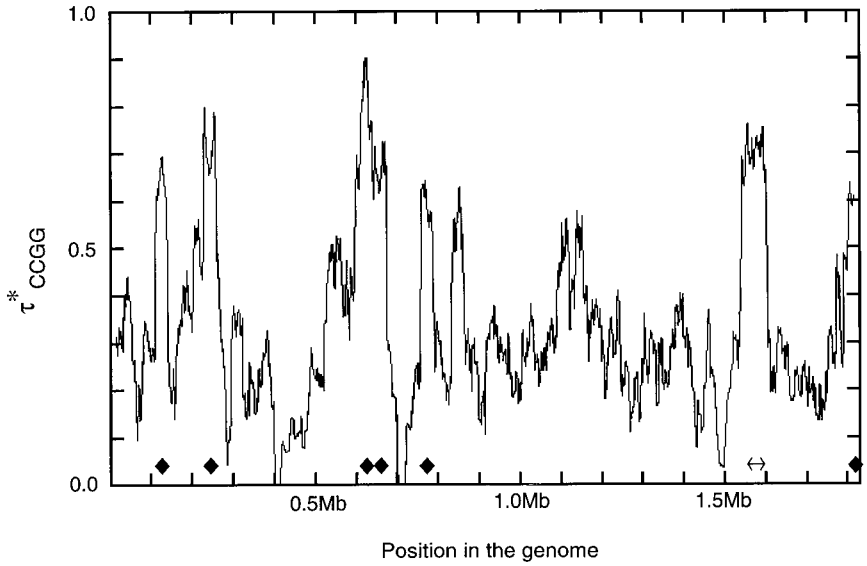

FIG. 1. τ^*_{CCGG} in a 30-kb sliding window along the *H. influenzae* genome. ◆, rRNA segments, ↔, the region with a preponderance of phage Mu-like sequences.

otides CGCG, GCGC, CTAG, TATA, and TCGA. In *E. coli*, CTAG sites peak in rRNA genes (33).

M. jannaschii. The *M. jannaschii* genome (1.66 Mb complete) features six significantly low palindromic tetranucleotides (Tables 3 and 4) and three high nonpalindromic tetranucleotides, each of the latter differing by a single base substitution from at least one of the low tetranucleotides. Roberts (55a), on the basis of sequence patterns, reports eight potential methylases of restriction modification systems. The counts and distributions of the palindromic tetranucleotides {CTAG, GATC, GTAC, CATG} of the same nucleotide content are striking. For example, CTAG occurrences are drastically low (total of 90), confined mainly to two significant clusters (by virtue of the *r*-scan statistics in Materials and Methods) 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 3,554 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. Apropos, a profile of G+C counts in 10-kb windows (or 50-kb windows) (Fig. 2) 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.

CTAG underrepresentations. CTAG is significantly underrepresented in many bacteria encompassing almost all purple proteobacteria, high-G+C gram-positive *Streptomyces*, and several archaeal genomes but generally not in eukaryotes (40). Although the tetranucleotide CTAG is very low in *E. coli* and *H. influenzae* (Table 3), the distribution of CTAG sites around the *E. coli* genome shows six significant clusters each contained in a rRNA unit (33), whereas in the *H. influenzae* genome, the *r*-scan statistics (see Materials and Methods) 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 at sharp variance to the mean frequency of CTAG in *E. coli* of about one per 5,200 bp over the whole genome. This anomaly applies to numerous other proteobacterial genomes. CTAG is generally low in most classes of *E. coli* phages (6). 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, perhaps as binding sites for regulatory proteins.

Except for *Streptomyces* genomes (e.g., *S. griseus*, *S. lividans*, and *S. coelicolor* [$\tau^* \leq 0.50$]), CTAG shows normal representations 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 *Mycobacterium tuberculosis*, *Mycobacterium leprae*, and *Clostridium glutamicum*. Moreover, CTAG possesses normal representations in all cyanobacterium sequences ($0.94 \leq \tau^*_{CTAG} \leq 1.04$) and is estimated in the normal to low normal range for all mycoplasmas (*M. genitalium*, $\tau^*_{CTAG} = 0.95$; *M. capricolum*, $\tau^*_{CTAG} = 0.83$), low normal in *Borrelia*

TABLE 4. Underrepresentation of palindromic tetranucleotides in various bacteria

Bacterium	No. of underrepresented palindromic tetranucleotides
Haemophilus influenzae	9
Methanococcus jannaschii	6
Neisseria gonorrhoeae	4
Helicobacter pylori	4
Streptococcus pneumoniae	4
Thermus sp.	4
Bordetella pertussis	3
Rhodobacter capsulatus	3
All other gram-positive bacteria	≤2
All cyanobacteria	≤2
Mycoplasma genitalium	1
Borrelia burgdorferi	2
Myxococcus xanthus	2

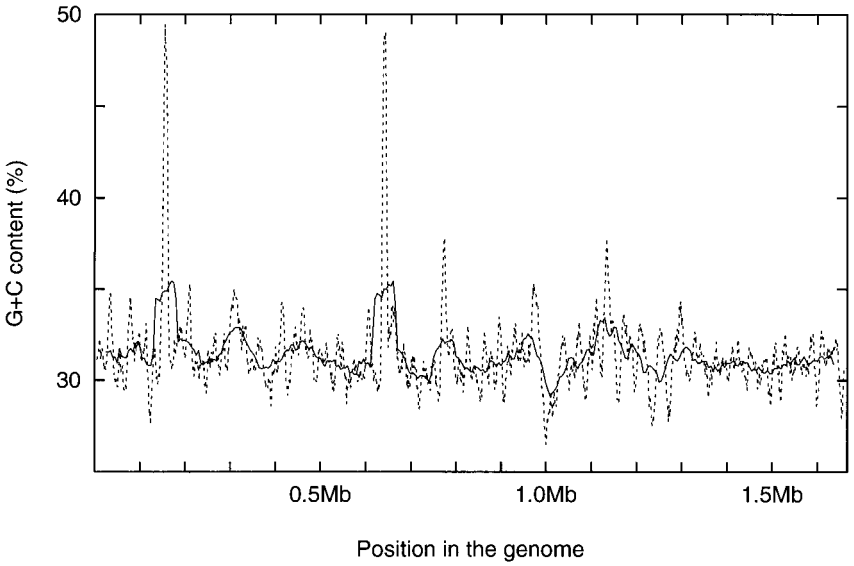

FIG. 2. G+C content in 50-kb sliding window (solid line) and 10-kb sliding window (dashed line) along the *M. jannaschii* genome.

burgdorferi ($\tau^* = 0.82$), and normal in *Chlamydia trachomatis* ($\tau^* = 0.96$).

Among archaea, CTAG is lowest in *M. jannaschii* and significantly low in *Halobacterium halobium* ($\tau^* = 0.55$) and in *M. thermoautotrophicum* ($\tau^* = 0.29$), but CTAG occurrences register in the normal range for *Sulfolobus* sp. sequences ($\tau^* = 0.96$) (34).

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.

Genome signature differences (δ* differences) within and between bacterial genomes. Large prokaryotic sequences having aggregate nonredundant DNA of about 200 kb or greater were compared via δ* differences (see Materials and Methods). Independent samples of about 100-kb lengths from each genome were formed. The ranges of δ* differences with respect to all samples within and between genomes are presented in Table 5. As an aid for our interpretations, Table 6 provides δ* differences for several familiar prokaryotic and eukaryotic examples (35, 36). Three hundred randomly generated sequences of 100-kb length show δ* differences persistently in the range 0 to 16 (all δ* differences henceforth are multiplied by 1,000). The average signature differences between 100-kb contigs within a single bacterial genome (diagonal of Table 5) range from 12 to 37, and those for 100-kb contigs between distinct genomes range from 37 to 267.

(i) δ* differences between proteobacterial sequences (Table 5). *E. coli* and *S. typhimurium* are close (average δ* = 37), about the same degree of similarity as among human chromosomes (40). *H. influenzae* is moderately similar (δ* differences of the order 50 to 75) to *E. coli*, both classified as γ-proteobacterial types and over twice as dissimilar from the α-proteobacterium *R. meliloti.*

(ii) Comparisons of gram-positive sequences. *B. subtilis* shows a level of weak similarity (δ* differences in the range 75 to 110) to the enteric γ-proteobacterial types and to the low-G+C-content gram-positive *S. aureus* and *Lactococcus lactis*. The latter genomic sequences are more different from *E. coli* and *S. typhimurium* sequences (δ* ≈ 92 to 129). *B. subtilis* compared to *E. coli* has average δ* difference 85, about the same as chicken to *Xenopus laevis* (Table 6).

(iii) Similarity comparisons of cyanobacterial genomes to proteobacterial and gram-positive sequences. A moderate similarity is observed between *Anabaena* sp. (not shown) and the gram-positive collections of *L. lactis* (δ* = 51) and *B. subtilis* (δ* = 56). The *Anabaena* and *Synechococcus* sequences are weakly similar (δ* = 86), whereas *Synechococcus* and *Synechocystis* sequences are dissimilar (δ* = 148). *Synechocystis* δ* differences to all proteobacterial and gram-positive sequences (except that of *L. lactis*) are pronounced, >150. Thus, by this measure, the three cyanobacterial (*Anabaena*, *Synechococcus*, and *Synechocystis*) genomes are not a coherent group. *Synechocystis* sp. sequences compared to proteobacterial sequences show at best weak similarities.

(iv) How dissimilar are the archaeal bacterial sequences from the eubacterial sequences? The average δ* differences of archaeal sequences from gram-negative proteobacterial sequences range from 137 to 248 (mostly ≥150) (Table 5). The greatest difference is generally to *M. jannaschii*. Differences of archaeal sequences from gram-positive sequences, though large, are not as extreme, δ* differences measuring in the range of 109 to 175 (mostly <150) (Table 5). These differences are consistent with protein sequence comparisons of heat shock proteins (HSP70), which place the archaeal HSP70 closer to gram-positive homologs than to gram-negative homologs (24–26).

The δ* differences between the thermophile *M. thermoautotrophicum* and other prokaryotes are comparable to those for *M. jannaschii*. Yet, the two thermophile archaea *M. jannaschii* and *M. thermoautotrophicum* have an average mutual δ* difference of 137. *H. halobium* is generally the most dissimilar (δ* ≥200) from all eubacterial sequences with one exception in that δ* differences from *Streptomyces* sequences are only about 80 (35, 39). *P. aerophilum* is substantially dissimilar from *M. jannaschii* and *M. thermoautotrophicum*, with δ* differences of 160 and 192, respectively, but tends to be, although distant, closer to classical bacteria (109 to 153) (Table 5).

The mutual δ* differences among the archaeal sequences place *M. thermautotrophicum*, *Sulfolobus* sp., and *M. jannaschii* at about 100 to 140, but differences from halobacterial se-

TABLE 6. Examples of dinucleotide relative abundance differences within and between disjoint eukaryotic and prokaryotic genomic collections based on ~100-kb disjoint sequence samples

Comparison	No. of samples	δ* (mean [range]) (10^3)
Random sequence[a]	300	9 (0–16)
Within *S. cerevisiae*	15	14 (3–29)
Within *E. coli*	46	20 (5–44)
Within human	14	35 (12–72)
Mouse vs rat	12 × 7	30 (11–58)
E. coli vs *S. typhimurium*	46 × 4	37 (17–60)
Human vs mouse	14 × 11	48 (16–94)
Human vs chicken	14 × 3	70 (50–92)
E. coli vs *B. subtilis*	46 × 5	85 (60–103)
Chicken vs *X. laevis*	3 × 3	88 (72–103)
Human vs sea urchin	14 × 2	106 (91–121)
Human vs *S. cerevisiae*	14 × 15	126 (103–151)
E. coli vs *M. genitalium*	14 × 6	156 (119–204)
Human vs *D. melanogaster*	14 × 9	177 (137–219)
Human vs *E. coli*	14 × 46	223 (187–262)
E. coli vs *Sulfolobus* sp.	46 × 1	231 (214–242)

[a] Random sequence of 100 kb with independently generated nucleotides. The δ* range is independent of the individual letter probabilities (e.g., G+C content).

quences are very large, about 250, suggesting a polyphyletic or highly diverse organization of the archaea (43, 55). The thermophilic genomes tend to be closer to vertebrate eukaryotes than to eubacterial sequences (35), whereas, as mentioned previously, the halobacterial sequences are weakly similar to *Streptomyces* gram-positive sequences.

(v) Partial-ordering comparisons of genomic sequences. We applied the partial-ordering method to large collections or complete genomes of sequences from *M. jannaschii*, *P. aerophilum*, *H. influenzae*, *M. genitalium*, *M. pneumoniae*, *E. coli*, *B. subtilis*, *Synechocystis*, maize, *Saccharomyces cerevisiae*, *Caenorhabditis elegans*, *Drosophila melanogaster*, *X. laevis*, chicken, mouse, and human. Figure 3 illustrates the resulting dominance relationships with the sequences of *M. jannaschii*, *M. pneumoniae*, *Synechocystis* sp., and human taken as standards. With respect to the *M. jannaschii* standard, the human, *X. laevis*, and *S. cerevisiae* sequences are among the undominated sequences, whereas the multiply dominated classical eubacterial sequences are farthest. As expected, the human sequence standard finds the other vertebrate sequences closest and undominated, with the classical eubacterial sequences farthest and substantially dominated. Interestingly, with the human standard, there are no partial orderings involving the *M. jannaschii* sequence, which is consistent with the fact that *M. jannaschii* genome is weakly similar ($\delta^* \approx 100$) to the human genome. The same is true for *Sulfolobus* (data not shown). With the *M. pneumoniae* standard, *M. genitalium* is the closest, whereas vertebrate and some archaeal sequences are among the most distant. Only few orderings emerge with the *Synechocystis* standard. In other words, most sequence collections are not comparable in terms of partial orderings, or equivalently, *Synechocystis* is an outgroup among currently available genomes.

Extensive iterations (microsatellites) in bacterial genomes. Table 7 reports the numbers of long microsatellite repeats (mono, di, tri, tetra) in the large bacterial sequence collections of Table 1. Strikingly, *E. coli* contains but a single nucleotide run of length ≥10 bp (explicitly G_{10}, intergenic) across its 4.6-Mb genome. There are 19 mononucleotide runs of length 9 bp where all but three are of type T_9 or A_9, mostly in association with terminator sequences. Also, *E. coli* shows but a single dinucleotide iteration of length 12 bp, $(GC)_6$ in an open reading frame (ORF) translated to $(Ala\ Arg)_2$. *H. influenzae*

TABLE 5. Differences between ~100-kb sequence samples (see Materials and Methods for details)

δ* differences (avg [range]) (10^3)														Comparison with:
γ-Proteobacteria			α-Proteobacteria	Low-G+C gram-positive bacteria			Mycoplasmas		Cyanobacteria		Archaea			
E. coli (46[a])	*S. typhimurium* (4)	*H. influenzae* (18)	*P. aeruginosa* (3)	*R. meliloti* (3)	*B. subtilis* (5)	*S. aureus* (3)	*L. lactis* (3)	*M. genitalium* (6)	*M. pneumoniae* (8)	*Synechococcus* (2)	*Synechocystis* (36)	*M. jannaschii* (17)	*M. thermoautotrophicum* (2)	*P. aerophilum* (22)	
20 (5–44)	37 (17–60)	57 (27–85)	91 (69–101)	141 (119–162)	85 (60–103)	92 (56–120)	118 (93–140)	156 (119–204)	150 (117–177)	95 (82–109)	152 (112–169)	228 (187–264)	199 (176–217)	151 (122–176)	*E. coli*
	21 (11–33)	73 (51–90)	103 (87–120)	124 (108–139)	89 (68–100)	102 (73–130)	129 (106–144)	179 (141–226)	162 (128–184)	108 (98–117)	152 (117–168)	217 (191–238)	221 (200–237)	139 (120–155)	*S. typhimurium*
		21 (6–44)	131 (112–149)	184 (168–200)	87 (74–100)	94 (68–122)	109 (76–129)	141 (102–199)	125 (89–156)	101 (91–111)	126 (91–148)	219 (181–252)	209 (188–228)	137 (106–164)	*H. influenzae*
			18 (14–25)	97 (86–108)	74 (59–89)	100 (91–113)	87 (78–93)	181 (155–218)	175 (153–191)	59 (46–70)	196 (155–213)	189 (170–207)	154 (147–164)	153 (129–174)	*P. aeruginosa*
				16 (13–20)	118 (103–131)	174 (155–199)	164 (154–172)	267 (237–307)	248 (226–269)	130 (119–137)	232 (186–247)	248 (234–261)	248 (235–262)	202 (183–220)	*R. meliloti*
					31 (19–51)	92 (61–125)	74 (45–100)	163 (122–220)	147 (111–179)	47 (34–61)	142 (97–160)	169 (134–196)	175 (155–194)	138 (103–176)	*B. subtilis*
						28 (27–29)	77 (61–93)	122 (93–164)	108 (92–134)	82 (71–93)	145 (109–167)	162 (136–193)	147 (133–155)	112 (88–135)	*S. aureus*
							22 (16–29)	108 (80–139)	105 (83–127)	51 (42–56)	127 (88–145)	126 (101–155)	118 (109–132)	109 (89–121)	*L. lactis*
								37 (16–71)	97 (64–135)	138 (116–173)	157 (123–183)	143 (117–167)	134 (128–142)	146 (127–188)	*M. genitalium*
									31 (16–54)	137 (120–153)	117 (84–142)	197 (158–231)	190 (171–212)	141 (110–173)	*M. pneumoniae*
										18	149 (108–167)	160 (141–176)	148 (143–152)	122 (114–133)	*Synechococcus*
											17 (3–48)	137 (103–170)	169 (142–185)	159 (121–195)	*Synechocystis*
												24 (6–51)	137 (112–153)	160 (129–181)	*M. jannaschii*
													12	192 (177–205)	*M. thermoautotrophicum*
														21 (5–44)	*P. aerophilum*

[a] Number of samples.

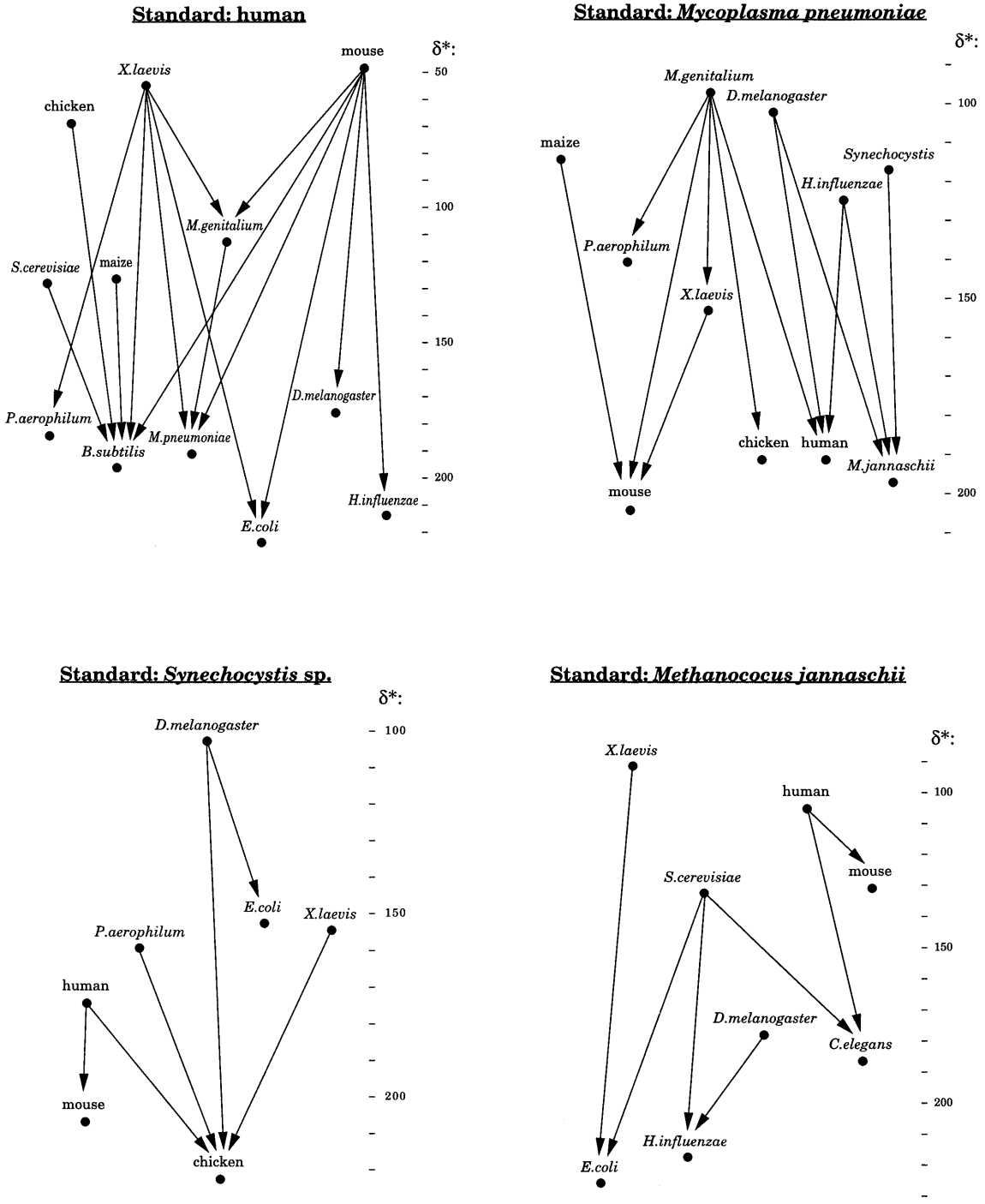

FIG. 3. Partial orderings with respect to human, *M. pneumoniae*, *Synechocystis* sp., and *M. jannaschii* sequence standards (see Materials and Methods for definition of partial orderings). Arrows indicate the partial ordering relationships (e.g., mouse is closer to the human standard than is *H. influenzae*). The vertical position of the species in the plots corresponds to the average δ* differences (100-kb sequence samples) to the standard (the scale is indicated on the right).

contains no dinucleotide iteration of length 12 bp or more. These results contrast sharply with the excessive numbers of microsatellites in the yeast genome, pervasive for each yeast chromosome (Table 8). Generally, microsatellites are frequent in eukaryotic genomes (data not shown) but rare in bacterial sequences except for trinucleotide repeats which obviously preserve the reading frame. For example, *M. genitalium* features 11 distinct long trinucleotide iterations, *Synechocystis* sp. has 14, and *E. coli* has 2. *M. genitalium* has five trinucleotide iterations of length ≥30 bp and an additional six iterations of ≥15 bp. Three of the five longest iterations are of the form $(AGT)_n$: $(AGT)_{16}$ at bp 169500 is intergenic, $(AGT)_{11}$ at bp 127150 appears in the gene *mgp* (MgPa operon) and translates to serine 11, and $(AGT)_{10}$ at bp 351450 is intergenic. The ORF MG338, 1,271 amino acids length, contains an iteration $(ACA)_{11}$ translated to threonine 11. An iteration $(CTT)_{16}$ im-

TABLE 7. Extended iterations poly(X), poly(XY), poly(XYZ), and poly(XYZW)

Sequence (length)	Count in whole genome (avg length, maximum length [bp])			
	poly(X), 10 bp[a]	poly(XY), 12 bp	poly(XYZ), 15 bp	poly(XYZW), 16 bp
E. coli (4.6 Mb)	1 (10, 10)	1 (12, 12)	2 (15, 15)	1 (18, 18)
H. influenzae (1.8 Mb)	2 (10, 10)	0	1 (29, 29)	12[b] (92, 148)
M. genitalium (580 kb)	1 (19, 19)	0	11[c] (32, 50)	1 (16, 16)
M. pneumoniae (816 kb)	3 (16, 16)	1 (22, 22)	2 (18, 21)	0
Synechocystis (3.6 Mb)	16 (10, 11)	0	14 (15, 17)	1 (18, 18)
M. jannaschii (1.7 Mb)	2[d] (16, 22)	0	1 (15, 15)	1 (17, 17)
P. aerophilum (2.2 Mb)	20 (12, 16)	3 (14, 14)	0	0
B. subtilis (508 kb)	0	0	0	0

[a] Minimum length of iterations. Long pentanucleotide iterations, i.e., poly(CTTCT) of ≥50 bp, are prominent in *Neisseria gonorrhoeae* and *Neisseria meningitidis.*

[b] Including $(CCAA)_{37}$, $(CCAA)_{21}$, $(CCAA)_{20}$, $(CCAA)_{19}$, $(CCAA)_{16}$, $(TCAA)_{33}$, $(TCAA)_{23}$, $(TCAA)_{17}$, $(GCAA)_{25}$, $(GACA)_{22}$, and $(AGTC)_{32}$.

[c] Including five iterations ≥30 bp in length.

[d] Including poly(G) 22 bp in length, 130 bp upstream of ORF MJ0312.

mediately followed by $(CTA)_8$ at bp 430000 is noncoding, lying between *recA* (MG339) and *rpo*C (MG340). It is also included in an MgPa repeat.

A formidable abundance of tetranucleotide microsatellites occur almost all in *H. influenzae* samples (11 times). Moxon et al. (49), for a number of pathogenic bacterial populations, highlight nonstandard mutation mechanisms which occur at special loci and explicitly discuss the case of the repeat tract $(TCAA)_{16}$ located in *H. influenzae* in the *lic2* gene near its 5′ end. This gene is essential for synthesis of digalactoside (27). In this circumstance, the loss or gain of one or more TCAA units may alter the reading frame and thereby regulate changes in the synthesis of digalactoside. More generally, tandem repeats (in the coding region and/or in gene regulatory regions) subject to homologous recombination or polymerase slippage during chromosomal replication can generate a heterogeneous population of cells (49). There are four occurrences of the tetranucleotide iterations $(TTGG)_{20}$, $(TTGG)_{20}$, $(TTGG)_{36}$, and $(CCAA)_{18}$ at kb 677, 706, 761, and 1633, respectively, whose flanking sequences extending about 2 kb downstream and about 400 bp upstream are substantially similar (~90% identical nucleotides) (41). As noted above, variation in the number of $(CCAA)_n$ iterations is a strategy that can alter the translational frame and/or intensity of DNA supercoiling in regulation of gene expression and can thus provide a bank of genetic polymorphism (41).

DISCUSSION

This report compares and contrasts genome-wide compositional biases and distributions of oligonucleotides across 15 diverse prokaryotic species that have substantial genomic sequence collections (including seven complete genomes). Our paramount observation pertains to the constancy within genomes of the dinucleotide relative abundance profile $\{\rho^*_{XY}$, all $XY\}$ (see Materials and Methods) over multiple disjoint 50-kb contigs. Also, the differences of the $\{\rho^*_{XY}\}$ vectors between genomes virtually always exceed those within genomes (in Table 5, compare off-diagonal with diagonal entries). This result prevails broadly for prokaryotes, eukaryotes, organelle, and viral DNA genomes, and accordingly, we refer to the $\{\rho^*_{XY}\}$ array as the genome signature (32, 39). On this basis, it appears generally that given the sequence of a 100-kb DNA contig, we can reasonably infer from its genomic signature to which group of organisms it belongs. Some caveats apply: closely related (by ancestry) species indeed tend to have similar genome signatures, while distantly related species have more dissimilar $\{\rho^*_{XY}\}$ profiles (6, 32, 36, 39). However, it is conceivable that in some cases, the genome signature differences could be small due to convergent evolution resulting from common ecological, physiological, and other selection forces.

The existence of genome-wide compositional biases raises questions of two types. (1) What molecular mechanisms and selective forces are responsible for these biases? (2) Can the signatures be useful as measures of phylogenetic relationships, and if so, what relationships do they indicate? Before discussing these questions, it is useful to highlight several of our findings.

Dinucleotide extremes. Many thermophiles (including *Thermus* sp., *M. thermoautotrophicum*, *M. jannaschii*, and *Sulfolobus* sp.) are significantly low in the dinucleotide CG. *Mycoplasma* sequences (e.g., *M. genitalium* and *M. capricolum*) but not *M. pneumoniae* are also low in CG. (*Mycoplasma* genomes are highly diverse and putatively derive in a polyphyletic manner [47] from various gram-positive origins). Other gram-positive sequences tend to have CG in the normal range, except for *Streptococcus* and *Clostridium* sequences, which are also low in CG. All γ-proteobacterial sequences show normal CG representations. On the other side, halobacterial sequences, sequences of several α-proteobacteria, and the *N. gonorrhoeae* sequence are significantly high in CG. For eukaryotes, CG suppression occurs in vertebrates, diverse protist genomes, dicot (but generally not monocot) plants, animal mitochondrial sets, and almost all vertebrate small viral genomes (Table 9 and reference 32). CG suppression in vertebrates has usually been ascribed to the classical methylation/deamination/mutation scenario causing mutation of CG to TG/CA. However, even where active CG methylases are present and increasing the mutation rate, it is not obvious to us that pure mutation pressure is the primary driving force. Certainly, this hypothesis cannot account for the pervasive CG suppression in animal mitochondria that lack the standard methylase activity and is

TABLE 8. Microsatellites in yeast

Chromosome (length)	Counts in whole chromosome (avg length, maximum length [bp])			
	poly(X), 10 bp[a]	poly(XY), 12 bp	poly(XYZ), 15 bp	poly(XYZW), 16 bp
I (227 kb)	62 (13, 36)	9 (17, 24)	12 (20, 30)	1 (16, 16)
II (807 kb)	178 (13, 31)	27 (18, 40)	25 (26, 65)	6 (18, 23)
III (315 kb)	65 (13, 23)	21 (18, 29)	11 (20, 33)	4 (18, 21)
IV (1,522 kb)	278 (12, 35)	54 (19, 37)	49 (20, 72)	7 (24, 54)
V (574 kb)	121 (13, 25)	17 (19, 36)	15 (20, 35)	2 (19, 24)
VI (270 kb)	56 (12, 21)	18 (19, 62)	6 (27, 46)	1 (18, 18)
VII (1,091 kb)	221 (13, 31)	45 (18, 35)	33 (20, 39)	5 (19, 22)
VIII (563 kb)	128 (13, 37)	22 (19, 36)	22 (19, 29)	4 (18, 18)
IX (440 kb)	89 (13, 29)	16 (16, 29)	15 (23, 59)	0
X (745 kb)	148 (12, 23)	26 (20, 41)	17 (20, 50)	4 (18, 22)
XI (666 kb)	128 (12, 31)	29 (17, 35)	25 (22, 41)	6 (22, 32)
XII (1,066 kb)	210 (13, 30)	43 (19, 64)	35 (18, 32)	4 (16, 17)
XIII (924 kb)	160 (13, 34)	31 (18, 32)	32 (24, 108)	4 (19, 26)
XIV (784 kb)	147 (13, 42)	29 (17, 32)	19 (18, 27)	5 (20, 27)
XV (1,091 kb)	199 (13, 28)	36 (17, 41)	31 (24, 63)	3 (16, 17)
XVI (948 kb)	170 (12, 27)	26 (18, 27)	47 (21, 48)	7 (18, 19)
Complete genome (12 Mb)	2,360 (13, 42)	449 (18, 64)	394 (21, 108)	63 (19, 54)

[a] Minimum length.

TABLE 9. Representative dinucleotide compositional extremes in bacteria and eukaryotes

Group	Compositional biases[a]								G+C content (%)
	TA	AT	CG	GC	CC/GG	AA/TT	CTAG	GATC	
Archaea									
M. jannaschii	0	0	−−−	0	++	0	−−−−	−−−	31
M. thermoautotrophicum	−	0	−−	0−	+	0	−−−	0	48
Sulfolobus sp.	0	0	−−	0	0+	0	0	0	36
H. halobium	−−	0	+	0	0−	0	−−	0	62
P. aerophilum	0	0	0	0	0	0	0	++	51
Eubacteria									
Gram-negative proteobacteria[b]	−, −−	0, ++[c]	0, +	+, 0	0	0, +	−−−[d]	0[e]	40–66
Rickettsia prowazekii	0	0	−	++	0	0	−	0	32
Clostridium sp.[f]	0	0	−−	+	+	0	0	0−	30
Streptococcus sp.[g]	−	0	−	0, ++	0	0	0	−−, 0	35–39
Streptomyces sp.[h]	−−−	0	0	0	0	0	−−−	0	70
All other gram-positive bacteria[i]	−−	0	0	0, +	0	0, +	0	0	35–60
Mycoplasma sp.	−	−	−−−	0	0	0, +	0	0	32
Eukaryotes									
Vertebrates	−	0	−−−	0	0, +	0	0	0	40–50
Echinoderms	−−	0	−−	0	0	0	0	0	45
Invertebrates (protostomes)	−	0	0	0, +	0	0, +	0, −	0	40–45
Fungi	−	0	0−	0	0	0	0	0	35–53
Protists	Mostly −	0	Mostly −	0, +	0, ++	0	0	0	24–55
Plants	−, 0−	0	−−, 0−	0	0	0	0	0	33–47
Organelles									
Protist mitochondria	0−	0, −−	0, −−	0, +	0, +++	0, ++	0	0	23–42
Animal mitochondria	0	0	−−	0	++	0	0	0, −	15–45
Chloroplasts	0	0	0	Mixed	+	0, +	0, −	0	26–39

[a] Symbols for overrepresentation: +, significant; ++, strong; +++, very strong. Symbols for normal range: 0−, low normal; 0, normal; 0+, high normal. Symbols for underrepresentation: −, significant; −−, strong; −−−, very strong. See also the footnote to Table 2. Combinations of symbols reflect differences among the group members. For example, 0, + indicates that most member species are in the normal range, while others are significantly high.

[b] Including α-proteobacteria *A. tumefaciens*, *R. leguminosarum*, *R. meliloti*, *B. japonicum*, *R. capsulatus*, and *R. sphaeroides*, β-proteobacteria *B. pertussis*, *A. eutrophus*, *Xanthomonas campestris*, *N. gonorrhoeae*, and *N. meningitidis*, and γ-proteobacteria *E. coli*, *S. typhimurium*, *Klebsiella pneumoniae*, *Serratia marcescens*, *Erwinia chrysanthemi*, *S. flexneri*, *Vibrio cholerae*, *Yersinia enterocolitica*, *A. pleuropneumoniae*, *H. influenzae*, *A. vinelandii*, *P. aeruginosa*, *P. fluorescens*, *P. putida*, *P. syringae*, and *A. calcoaceticus.*

[c] γ-proteobacteria and *Neisseria* are normal; α-proteobacteria and other β-proteobacteria are ++.

[d] Some gram-negative proteobacteria are − or 0−; *H. pylori*, *N. gonorrhoeae*, and *V. cholerae* are normal.

[e] Excepting *Neisseria* −−.

[f] *C. acetobotulicum* and *C. perfringens.*

[g] *S. pneumoniae* and *S. mutans.*

[h] *S. coelicolor*, *S. griseus*, and *S. lividans.*

[i] *B. subtilis*, *S. aureus*, *B. brevis*, *B. stearothermophilus*, *Coxiella burnetti*, *L. lactis*, *M. leprae*, *M. tuberculosis*, and *C. glutamicum.*

unlikely to apply to most of the bacteria analyzed herein. Moreover, some mammalian genomes and all animal mitochondrial genomes have CC/GG high but TG/CA in the normal range, portending a CG→CC/GG mutation bias. We have proposed that CG deficiencies may impart a selective advantage due to structural constraints related to high dinucleotide stacking energy, supercoiling, or chromatin packing (32).

TA is significantly underrepresented across most prokaryotic and eukaryotic genomes. However, *P. aerophilum*, like *Sulfolobus*, is strictly normal in TA representations. Likewise, TA is normal, $\rho^*_{TA} \approx 0.98$ to 1.03 in almost all mitochondrial and chloroplast genomes (34). Possible contributing influences to the widespread underrepresentation of TA are the following: (i) TA has the least thermodynamically stable DNA duplex of all dinucleotides (8, 16), entailing flexibility of the TA site for unwinding the DNA double helix; (ii) RNases preferentially degrade UpA dinucleotides in mRNA tracts (3); and (iii) TA is part of many regulatory sequences (e.g., TATA box and termination signals) so that reduced TA usage may help avoid inappropriate binding of regulatory factors. Untwisting and bending at TA sites occurs in much of transcription initiation via protein binding, for example, to the TATA box, *Eco*RV binding to its recognition sequence, and γδ resolvase binding to the site at which crossing-over occurs (61). These models suggest that TA sites can be important as nucleation sites for untwisting the DNA double helix.

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

CTAG is significantly low in virtually all gram-negative purple proteobacterial sequences but of normal representations in cyanobacterial sequences and in gram-positive genomes (excepting *Streptomyces* sequences). Archaeal sequences are quite variable 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 (34). The *M. jannaschii* genome is unprecedented in the extremely low relative abundance value of its CTAG tetranucleotides. Specifically, 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 CTAG occur in the region from 636994 to 643016.

Interpretations of underrepresentations of CTAG center on structural defects (kinking) or special functional roles associated with this tetranucleotide (11, 23). In this context, the crystallographic resolutions of the TrpR-DNA complex (52) and also for the MetJ-DNA complex (54) indicate CTAG kinks which 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 (4, 33).

Tetranucleotide biases in eukaryotes are relatively uncommon (18 of the 33 genomes with substantial DNA available [40] show no tetranucleotide over- or underrepresentations). Palindromic tetranucleotides, unlike in bacterial genomes, are in a number of cases overrepresented in higher eukaryotes (40). More specifically, CGCG carries high relative abundances in most vertebrates, in dicot plants, and in the yeast species *Kluyveromyces lactis* and *Candida albicans*. Notably, all of these entail significant CG suppression.

Genomic comparisons (δ^* differences and partial orderings). *Synechocystis* deviates substantially (δ^* differences of $\geq$150 [Table 5]) from the cyanobacterial *Synechococcus* and *Anabaena* sp. sequences. In this context, the three major classes of cyanobacteria do not constitute a coherent group and are generally as far from each other as are gram-negative from gram-positive sequences (Table 5). Moreover, gram-negative and gram-positive bacteria are themselves highly diverse clades.

δ^* differences of *M. jannaschii* sequence to all gram-negative proteobacterial sequences are very large, $\geq$180. The corresponding differences between *M. jannaschii* and low-G+C Gram-positive bacteria are high, in the range 130 to 160, which is about 30% closer (especially to *B. subtilis*, *L. lactis*, and *S. aureus*). This analysis placing the thermophile archaea (also the halophile archaea [35]) much closer in δ^* differences to gram-positive sequences than to gram-negative sequences is in agreement with the Gupta and Golding (25) assessments of bacterial sequence similarities, based on heat shock protein (HSP70) sequence comparisons.

In δ^* differences *M. jannaschii* (and *Sulfolobus*) are more similar by a factor of 2 or 3 to eukaryotes (especially human and yeast [Fig. 3]) than to gram-negative proteobacteria.

Special genome features. *Synechocystis* sp. is different from the other genomes with respect to (i) overrepresentation of all homodinucleotides and high numbers of long homonucleotide runs and (ii) the very frequent 10-bp palindrome (GGCGAT CGCC) (41). Their high density and significantly even distribution around the genome suggest that they may contribute to genome-wide activities such as replication and repair, sites of membrane attachments in association with domain loops, sites of nucleating Okazaki fragments or helix unwinding, and/or sites contributing to genome packaging. Longer palindromes are scarce. Close dyads are relatively rare in *Synechocystis* compared to the other bacterial sequences under study (data not shown).

In our prokaryotic compositional analysis, *H. influenzae* stands out in two ways: (i) an impressive number (nine) of underrepresented palindromic tetranucleotides (Table 3) and concomitantly many extant restriction systems and (ii) the preponderance of long tetranucleotide iterates (Table 7), many in coding regions, virtually absent from the other prokaryotic sequences under study. *H. influenzae* is also distinguished by the multitude of USSs (uptake signal sequences) in the genome vital for successful incorporation of heterologous DNA into the *Haemophilus* genome, where the absorbed sequence requires copies of USS in it. Putatively, the densely spread USS motif coupled to the abundant restriction system repertoire of *Haemophilus* genome provide barriers to lateral gene transfer of foreign DNA. The significantly even spacings of USS sites (41) may be essential in replication and repair processes such that heterologous sequences lacking enough USS placements generate impaired genomes.

Possible mechanisms underlying the genome signature. The discrimination between genomes of prokaryotes and eukaryotes that is afforded by δ^* differences is significantly robust although the underlying mechanisms are hardly understood. Dinucleotide relative abundances capture most of the departure from randomness in DNA 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 (35, 36), suggesting that DNA conformational stacking arrangements are principally determined by base-step configurations (8, 16, 28). Observation of the distribution of dinucleotide relative abundances separated by no, one, or two other nucleotides has shown that although values for no separation are often highly biased, those for separation by one or two nucleotides are more nearly random (32).

The fact that the dinucleotide signature pervades the entire genome leads us to attribute it to some genome-wide process(es), specifically to replication and/or repair. The signature might relate to replication in two basically different ways: (i) the replication/repair machinery might generate context-dependent mutation rates (as in the conventional explanation for CG suppression as a consequence of cytosine methylation) or (ii) the replication apparatus (including not just DNA chain elongation but the attendant requirements for chromosomal segregation and function) might operate more efficiently on specific sequences than on others. In the first case, certain dinucleotides are preferentially generated; in the second, they are selected through their effects on cellular phenotype.

We hypothesize that differences between organisms in replication and repair machinery largely maintain the homogeneity of the whole genome of an organism and that this is reflected in the genome signature. We indicate a suggestive example. The dinucleotide relative abundance values of temperate double-stranded DNA phages are very close to their hosts, filamentous and single-stranded DNA phages are moderately to distantly related to their hosts, and lytic double-stranded DNA phages are generally distant from their hosts, with phage T7 being substantially farther than phage T4 (6). This gradient in similarity to the host parallels the extent to which the phage uses the complete replication and repair machinery of the host and the duration of such use (6).

DNA structural configurations appear to be largely determined by base-step (double-strand) dinucleotide arrangements (8, 12, 16, 28, 32, 61). Hunter (28) set forth a theoretical framework for understanding and predicting the sequence-dependent structure and properties of double-stranded DNA. The analysis derives primarily from the energetics of base stacking interactions. These take account of cross-strand steric clashes (for example, at pyrimidine-purine steps) and of electrostatic interactions between partial atomic charges and the π

electrons of the DNA nucleotide aromatic rings. Furthermore, a study of the energy minima for the geometry of two neighboring base pairs in terms of slide, roll, and helical twist parameters finds that the 16 possible base steps can largely account for the DNA structures of synthesized oligonucleotides (12, 28) determined by X-ray diffraction. The structure of longer oligonucleotides to a significant extent can be predicted on the basis of dinucleotide base step interactions (12, 28, 32, 61). Phillips et al. (53) have concluded from a Markov chain study of *E. coli* sequences that "constraints affecting oligonucleotide frequencies occur at the trinucleotide level or lower."

Dinucleotide relative abundance variation putatively reflect duplex curvature, supercoiling, and other higher-order DNA structural features. Many DNA repair enzymes putatively recognize shapes or lesions in DNA secondary structures more than specific sequences (19, 42). Nucleosome positioning, interactions with DNA-binding proteins, and ribosomal binding of mRNA are strongly affected by dinucleotide arrangements (61, 66). Certain base steps are associated with an intrinsic curvature, which can lead to bending and supercoiling. DNA structures may be crucial in modulating processes of replication and repair.

Other general factors influencing DNA structure include exposure to sunlight (effects of UV irradiation), osmolarity (e.g., salt concentrations), hydrostatic pressure, acidity and alkalinity tolerance, extreme temperature, and alcohol ambience. There appear to be nucleotide biases in replication, in mutagenesis, and in rates of insertions and deletions dependent on neighboring base context (42). Stacking capacities may influence base incorporation rates and choices.

Genomic flux. Prokaryotic genomes are in a dynamic condition influenced by natural genetic transformation (competence), transposition, recombination, inversion, duplication, deletion, and possible fusion events. Substantial mixing of DNA material from diverse sources, a priori, seems in conflict with the constancy of the genomic signature profile $\{\rho^*_{XY}\}$. Nevertheless, the data strongly support the validity of the genomic signature. For resolution of this conundrum, see below.

Nearly all cells of *H. influenzae* and *N. gonorrhoeae* are competent (11, 46, 59). Only a small percentage (~10%) of *B. subtilis* cells appear to be competent for uptake of nonspecific DNA sequences (46, 59). Specifically, in *B. subtilis* and *Streptococcus pneumoniae*, competence appears to be regulated by cell density, cell-cell communication, and nutritional signaling dependence on growth conditions (59). Therefore, in this case, DNA uptake is likely to be mostly of similar DNA. Generally, although exogenous DNA incorporation is widespread in bacterial cells, nonspecific integration into the chromosome seems to be rare (46).

A major hypothesis concerning *H. influenzae* (and some other bacterial organisms) is that natural genetic competence (transformation) evolved and is maintained for the purpose of acquiring templates mediating repair of DNA lesions (41, 48). Other possible roles of natural genetic competence are benefits of horizontal gene transfer (e.g., transfer of antibiotic resistance determinants), repair of damaged chromosomes (that are rescued by recombination with exogenous homologous DNA), conversion of mutant alleles to functional alleles, or simply furnishing a good nutrient source (46, 48). Natural genetic transformation among bacteria generally accepts DNA of a conspecific strain but rarely of an exotic species. We further speculate that DNA acquired by horizontal gene transfer is rapidly converted to the genome signature of its new host. Perhaps the simplest argument for rapid acquisition comes from studies of many bacteriophages, whose genomes, ostensibly primarily chimeric in origin, exhibit uniform signatures (6).

The biochemical nearest-neighbor analyses (29, 57, 58) might be used to investigate the effects of altered replication and repair factors and context-dependent mutational tendencies. For example, in *E. coli*, DNA polymerases I, II, and III, and their associated factors and appropriate control elements, might be replaced by those of *B. subtilis* or some other weakly similar bacterium followed by tracking time changes in the dinucleotide relative abundance genome signature. Without defending the practicality of carrying out such an experiment in real time, we consider it worth mentioning because it defines our perspective on the principal explanation for genomic signatures. Because the signature pervades the entire genome, a natural way to explain it relates to repair and replication processes. Rapid (within 1,000 generations) significant change in global G+C content has been observed in the mutator strain of *E. coli* (15). It would be of interest to evaluate dinucleotide relative abundances in the mutator strain of *E. coli* during its process of change.

Genome signature and phylogeny. At least three chronometers have been applied in appraising similarities and dissimilarities among various genomes.

First, with respect to the original 16S rRNA comparisons, the validity of rRNA comparisons has been argued as follows (51): these genes are (i) present in all cellular genomes, (ii) conservative in their rates of change, and (iii) unlikely to be exchanged among lineages by horizontal gene transfer. These genes contain limited information. They span only about 1,500 to 1,800 nucleotides, of which only about half are ordinarily retained in attempting to develop informative alignments.

Second, protein sequence comparisons likewise require alignable segments. The amount of sequence available for comparison for the ensemble of all proteins is much greater than that of 16S rRNA. The results of such analyses are mixed and conflicting (see section below).

Third, genomic signature comparisons (δ^* distances and partial orderings) utilize sequence information from entire genomes (coding and noncoding) with no requirement for alignment.

Conventional methods of phylogenetic reconstruction from sequence information (the first and second methods noted above) use only similarity or dissimilarity assessments of aligned homologous genes or regions (20, 44, 45, 50). Difficulties intrinsic to these methods include the following: (i) alignments of distantly related long sequences (e.g., complete genomes) are generally not feasible; (ii) different phylogenetic reconstructions (trees) may result for the same set of organisms based on analysis of different protein, gene, or noncoding sequences (attempts are made to overcome this by averaging over many proteins [18a]); (iii) resultant trees may be highly dependent on details of the alignment algorithm used; (iv) the often made assumption of constant rates of evolution on the various branches of the tree or at different sites within a sequence may be violated (the problem of unequal rate effects [44]); (v) chimeric origins, recombination, inversions, transpositions, and lateral transfer between distantly related organisms may complicate analyses; and (vi) tree construction derived from aligned sequences cannot apply to organisms for which similar gene sequences are largely unavailable (e.g., for bacteriophages, eukaryotic viruses, or deeply divergent organisms [6, 35]).

The analysis of dinucleotide relative abundance values for phylogenetic analyses has the following advantages: (i) it does not depend on finding homologous genes in the sequences compared; (ii) it does not require a prior alignment and is

unaffected by the presence of gaps and large rearrangements in the sequence; (iii) the genomic relative abundance differences can use the entire available genome sequence data for the organisms; and (iv) for 50-kb (or longer) contig samples, the genome signature has remarkably small variance such that the average δ^* differences for multiple samples of 50-kb contigs between genomes almost always substantially exceed within-genome differences.

The genomic signatures of δ^* difference among vertebrates imply orderings consistent with accepted phylogenetic reconstructions (36, 40). Similarly, the δ^* differences among major fungal sequences are consonant with accepted orderings (36). Genome signature comparisons have been applied to a wide assortment of bacteriophage genomes (6). We refer to reference 39 for results on δ^* differences applied to more than 40 prokaryotic sequences, each having at least 100 kb total of nonredundant genomic sequences. However, translation of sequence similarities into evolutionary relatedness will always be tentative, as the underlying assumptions about mutation rates, selective forces, and gene transfer events are uncertain.

Domains of life and the origin and early evolution of eukaryotes. Our discussion has centered heretofore on the role of the genome signature in highlighting similarities and dissimilarities across different classes of prokaryotic species. Related discussion concerned possible mechanisms underlying the genome signature, the extent and nature of the genome compositional flux, and the use of the genome signature as a chronometer for molecular phylogeny. Here we consider possible implications of the genome signature relative to current hypotheses for the major kingdoms (domains) of life and the genesis of organelles. Most seminal ideas in science are generated early in the development of a field, when the available facts are limited. As more information becomes available, there is a tendency to explain new facts by employing familiar hypotheses rather than to reassess the entire conceptual framework. Before suggesting alternatives, it is useful to review briefly the current proposals on domains of life.

Relevant to the origin of eukaryotes, we cite Woese's main observation (63, 65), based on 16S rRNA comparisons, that three separate domains of life could be distinguished: eubacteria (abbreviated B), archaea (A), and eukaryotes (K). With respect to their early evolution, an analysis again based initially on 16S rRNA was that the genomes of major organelles (mitochondria and chloroplasts) are more closely akin to those of eubacteria than to the nuclear genomes of eukaryotes. This fact was taken to support the endosymbiont hypothesis, that such organellar genomes constitute the remnants of once free-living intracellular parasites.

It is crucial to specify in what sense these data support the endosymbiont hypothesis. The endosymbiont hypothesis has been proposed repeatedly throughout this century and has many attractive features. The alternative is that the organellar genome is composed of genes derived from the nuclear genome that became sequestered in the organelle. The dissimilarity between the nuclear genome and organellar genomes speaks against a nuclear origin. We will return to this question later.

Both the three domain hypothesis and the endosymbiont hypothesis have undergone subsequent refinement. First, the original reason for dividing the living world into three and only three domains was that there were, on the initial evidence, only three approximately coherent sets and that these were about equally distant from one another. Insistence on three domains leaves frozen a classification based on the limited knowledge available in the past. If A and K are more closely related than either is to B (another point of controversy), then A and K are in the same domain. Otherwise, why not proceed further to define additional primary domains by splitting at the first nodes within any of these three domains? Rivera and Lake (55), among others, suggest four domains: eubacteria, halophiles, eocytes (hyperthermophilic sulfur-metabolizing bacteria, e.g., *Sulfolobus*), and eukaryotes. In another analysis, Woese, Pace, and collaborators (summarized in reference 64) recognize a deep split of the archaea into *Crenarchaeotes* and *Euryarchaeotes* and lately another subgroup, Korarchaeota (see also reference 18). The tripartite (eubacteria, archaea, prokaryotes) description and monophyletic nature of the archaea are under strong debate (1, 25, 26, 39, 40, 43, 62).

The endosymbiont hypothesis has been refined in two major ways. First, as it became increasingly apparent (especially for mitochondria) that many organellar functions are encoded in the nucleus, it was assumed that these nuclear genes had been relocated to the nucleus by lateral transfer from the organellar genome. (A reason why some genes have remained in the organelle is then needed.) Second, 16S rRNA phylogenies also required that, at least for chloroplasts, existing organelles are descended not from one endosymbiont but from several which invaded different lineages at different times (21). A central unresolved problem concerns whether mitochondrial evolution is monophyletic or polyphyletic (21, 22). Moreover, there is substantial evidence for secondary loss of the mitochondrion from various protists (e.g., *Entamoeba histolytica*) in which several well identified mitochondrial genes are found in the nuclear chromosomes (14). Multiple independent losses and gains of genes (and of full mitochondria and chloroplasts) is probably the norm. Many metabolic systems, including the ancestral respiratory system, were lost in diverse evolutionary lines in response to adaptation to different niches (13).

The extensive sequence information now available on various genomes provides a much richer basis for appraising similarities among them. As mentioned earlier, at least three complementary approaches have been taken. (i) 16S rRNA comparisons place A and K closer to each other than to B. (ii) Protein sequence comparisons give mixed and conflicting results (1, 2, 9, 24, 26, 50, 55) [for example: (a) EF-1a, (EF-2G) {A, K}; (b) Rad51/Dmc1/RadA/RecA→{A, K}; (c) RNA polymerase, A and C subunits→{A, K}; (d) HSP70→{A, B}; (e) glutamine synthetase→{A, B}; and (f) glutamate dehydrogenase→{A, B}].

Most of these relations are inferred from analysis of protein families possessing a multiple sequence alignment that commonly reveals a signature amino acid segment (or module) present in or absent from sequences of appropriate subgroups of the family. For example, with respect to the RecA-like sequences, the A module (7) separates {A, K} from {B}. But with respect to the HSP70/DnaK protein family, there is an amino acid segment absent from {A and gram-positive B} but present in {K} and most gram-negative proteobacterial sequences (26). For HSP70, the archaea (especially certain halobacterial and methanococcal sequences) are closer to selected gram-positive sequences (26, 30).

To reconcile the conflicting protein sequence analyses, Gupta and Golding (25) propose a chimeric eukaryotic cell nucleus resulting from fusion of an eocyte (*Sulfolobus*-like) bacterium and a gram-negative bacterium, to form a "unique" chromosomal transformation preceding mitochondrial and chloroplast endosymbiotic events. Other authors emphasize lateral gene transfer as a major mechanism for interpreting multiple alignments. This applies to the discussions of glutamine synthetase (9, 56, 60) and of glutamate dehydrogenase (2). For interpretations of EF-1α prokaryotic/eukaryotic sequence relationships, see references 1 and 55.

In sequence comparisons of genes from *M. jannaschii* to those of other genomes, it is observed that some genes are totally bacterium-like, while others distinctly resemble eukaryotes (10, 18). It is further noted that the most significant matches of many genes of *M. jannaschii* are to genes from human and yeast (10). This agrees with our global genome signature comparisons with respect to both partial orderings and δ^* differences (Fig. 3), which place *M. jannaschii* much closer to yeast and human than to the classical eubacteria.

At the metabolic level and with respect to transport across the cell membrane, many archaeal thermophiles and classical bacteria appear to have their central biochemical pathways derived from a common ancestor (18). On the other hand, with respect to genes important in information processing systems (replication, transcription, and translation), the thermophilic archaea are more similar to eukaryotes. Thus, eukaryotic nuclear gene sequences seem to be of two types: metabolic housekeeping proteins that are mostly related to eubacterial counterparts and proteins of information systems that are mostly related to those in some archaeal genomes.

Genomic signature comparisons (δ^* distances and partial orderings) favor the association {A, K} for several thermophiles of A but the association {A, B} for various halophiles (35, 39). Table 9 juxtaposes the most outstanding dinucleotide relative abundance values for representatives of the {B}, {A}, and {K} domains.

None of these results is incompatible with the endosymbiont hypothesis, but they do undermine its evidential basis. If the primordial eukaryotic nucleus was already a mosaic of genes from A and B (due either to a single nuclear fusion, as suggested by Gupta and Golding (25), or to extensive lateral transfer between eubacteria and archaea), then a nuclear origin of mitochondria is no longer excludable. In the initial argument, 16S rRNA was used as the sole chronometer, and so the entire genome of nucleus or organelle was assumed to have a common origin. With the range of protein sequences now available, the nuclear genome clearly appears to be chimeric. If it arose by fusion of two entire genomes, as Gupta and Golding (25) propose, each genome must have had its own 16S rRNA, one of which might then have broken off to inhabit the organelle. We conclude that the virtues of the endosymbiont hypothesis must be argued on some basis other than molecular phylogenies that are based on sequence alignments. Another basis could be phylogenies based on conservation of genomic signatures.

This leaves many possible scenarios. Our personal favorite compresses the Gupta-Golding fusion and the endosymbiont invasion into a single event. The chimeric nature of the nuclear genome could then result primarily from migration into the nucleus of many genes, not just those affecting organellar function. We consider the *Sulfolobus* lines as a likely candidate for the endosymbiont, particularly of animal mitochondria, for reasons previously outlined (34). These reasons include (but are by no means restricted to) similarities in genome signature. As indicated earlier, the uniformity of the signature throughout each genome suggests its rapid acquisition (on an evolutionary time scale). Therefore, if a considerable time period had elapsed between a cell fusion event and organelle formation, one might expect that any genes that later migrated into the organelle might have lost their original characteristic signature. Although this argument embodies many untested assumptions, it is the only one presently defensible that uses genome sequences to favor an endosymbiotic origin of organelles. Progress in our understanding of genomic evolution and phylogenetic relationships may require synthesis of sometimes conflicting results from rRNA, protein, and genome signature comparisons.

ACKNOWLEDGMENTS

S.K. was supported in part by NIH grants 5R01GM10452-32 and 5R01HG00335-08 and NSF grant DMS9403553-002. A.M.C. was supported in part by grant 9R01GM51117-28.

REFERENCES

1. **Baldauf, S. L., J. D. Palmer, and W. F. Doolittle.** 1996. The root of the universal tree and the origin of eukaryotes based on elongation factor phylogeny. Proc. Natl. Acad. Sci. USA **93:**7749–7754.
2. **Benachenhou-Lahfa, N., P. Forterre, and B. Labedan.** 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:**335–346.
3. **Beutler, E., T. Gelbart, J. Han, J. A. Koziol, and B. Beutler.** 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:**192–196.
4. **Bhagwat, A. S., and M. McClelland.** 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:**1663–1668.
5. **Bishop, Y. M. M., S. E. Fienberg, and P. W. Holland.** 1975. Discrete multivariate analysis: theory and practice. MIT Press, Cambridge, Mass.
6. **Blaisdell, B. E., A. M. Campbell, and S. Karlin.** 1996. Similarities and dissimilarities of phage genomes. Proc. Natl. Acad. Sci. USA **93:**5854–5859.
7. **Brendel, V., L. Brocchieri, S. J. Sandler, A. J. Clark, and S. Karlin.** 1997. Evolutionary comparisons of RecA-like proteins across all major kingdoms of living organisms. J. Mol. Evol. **44:**528–541.
8. **Breslauer, K. J., E. Frank, H. Blöcker, and L. A. Marky.** 1986. Predicting DNA duplex stability from the base sequence. Proc. Natl. Acad. Sci. USA **83:**3746–3750.
9. **Brown, J. R., Y. Masuchi, F. T. Robb, and W. F. Doolittle.** 1994. Evolutionary relationships of bacterial and archaeal glutamine synthetase genes. J. Mol. Evol. **38:**566–576.
10. **Bult, C. J., O. White, G. J. Olsen, L. Zhou, R. D. Fleischmann, G. G. Sutton, J. A. Blake, L. M. FitzGerald, R. A. Clayton, J. D. Gocayne, et al.** 1996. Complete genome sequence of the methanogenic archaeon, *Methanococcus jannaschii*. Science **273:**1058–1073.
11. **Burge, C., A. M. Campbell, and S. Karlin.** 1992. Over- and under-representation of short oligonucleotides in DNA sequences. Proc. Natl. Acad. Sci. USA **89:**1358–1362.
12. **Calladine, C. R., and H. R. Drew.** 1992. Understanding DNA. Academic Press, San Diego, Calif.
13. **Castresana, J., and M. Saraste.** 1995. Evolution of energetic metabolism: the respiration-early hypothesis. Trends Biochem. Sci. **20:**443–448.
14. **Clark, C. G., and A. J. Roger.** 1995. Direct evidence for secondary loss of mitochondria in *Entamoeba histolytica*. Proc. Natl. Acad. Sci. USA **92:**6518–6521.
15. **Cox, E. C., and C. Yanofsky.** 1967. Altered base ratios in the DNA of an *Escherichia coli* mutator strain. Proc. Natl. Acad. Sci. USA **58:**1895–1902.
16. **Delcourt, S. G., and R. D. Blake.** 1991. Stacking energies in DNA. J. Biol. Chem. **266:**15160–15169.
17. **Dembo, A., and S. Karlin.** 1992. Poisson approximations for r-scan processes. Ann. Appl. Prob. **2:**329–357.
18. **Doolittle, W. F.** 1996. At the core of the Archaea. Proc. Natl. Acad. Sci. USA **93:**8797–8799.

18a. **Doolittle, R. F., D.-F. Feng, S. Tsang, G. Cho, and E. Little.** 1996. Determining divergence times of the major kingdoms of living organisms with a protein clock. Science **271:**470–477.

19. **Echols, H., and M. F. Goodman.** 1991. Fidelity mechanisms in DNA replication. Annu. Rev. Biochem. **60:**477–512.
20. **Felsenstein, J.** 1988. Phylogenies from molecular sequences: inference and reliability. Annu. Rev. Genet. **22:**521–565.
21. **Gray, M. W.** 1992. The endosymbiont hypothesis revisited. Int. Rev. Cytol. **141:**233–357.
22. **Gray, M. W.** 1995. Mitochondrial evolution, p. 635–659. *In* C. S. Levings III and I. K. Vasil (ed.), The molecular biology of plant mitochondria. Kluver Academic Publishers, Dordrecht, The Netherlands.
23. **Gunsalus, R. P., and C. Yanofsky.** 1980. Nucleotide sequence and expression of *Escherichia coli* trpR, the structural gene for the trp aporepressor. Proc. Natl. Acad. Sci. USA **77:**7117–7121.
24. **Gupta, R. S., K. Bustard, M. Falah, and D. Singh.** 1997. Sequencing of heat shock protein 70 (DnaK) homologs from *Deinococcus proteolyticus* and *Thermomicrobium roseum* and their integration in a protein-based phylogeny of prokaryotes. J. Bacteriol. **179:**345–357.
25. **Gupta, R. S., and G. B. Golding.** 1996. The origin of the eukaryotic cell. Trends Biochem. Sci. **21:**166–171.
26. **Gupta, R. S., and B. Singh.** 1994. Phylogenetic analysis of 70 kD heat shock

protein sequences suggests a chimeric origin for the eukaryotic cell nucleus. Curr. Biol. **4:**1104–1114.
27. **High, N. J., M. E. Deadman, and E. R. Moxon.** 1993. The role of a repetitive DNA motif (5′-CAAT-3′) in the variable expression of the *Haemophilus influenzae* lipopolysaccharide epitope αGal(1-4)βGal. Mol. Microbiol. **9:** 1275–1282.
28. **Hunter, C. A.** 1993. Sequence-dependent DNA structure: the role of base stacking interactions. J. Mol. Biol. **230:**1025–1054.
29. **Josse, J., A. D. Kaiser, and A. Kornberg.** 1961. Enzymatic synthesis of deoxyribonucleic acid. VIII. Frequencies of nearest neighbor base sequences in deoxyribonucleic acid. J. Biol. Chem. **236:**864–875.
30. **Karlin, S.** 1995. Statistical significance of sequence patterns in proteins. Curr. Opin. Struct. Biol. **5:**360–371.
31. **Karlin, S., and V. Brendel.** 1992. Chance and statistical significance in protein and DNA sequence analysis. Science **257:**39–49.
32. **Karlin, S., and C. Burge.** 1995. Dinucleotide relative abundance extremes: a genomic signature. Trends Genet. **11:**283–290.
33. **Karlin, S., C. Burge, and A. M. Campbell.** 1992. Statistical analyses of counts and distributions of restriction sites in DNA sequences. Nucleic Acids Res. **20:**1363–1370.
34. **Karlin, S., and A. M. Campbell.** 1994. Which bacterium is the ancestor of the animal mitochondrial genome? Proc. Natl. Acad. Sci. USA **91:**12842–12846.
35. **Karlin, S., and L. Cardon.** 1994. Computational DNA sequence analysis. Annu. Rev. Microbiol. **48:**619–654.
36. **Karlin, S., and I. Ladunga.** 1994. Comparisons of eukaryotic genomic sequences. Proc. Natl. Acad. Sci. USA **91:**12832–12836.
37. **Karlin, S., and C. Macken.** 1991. Assessment of inhomogeneities in an Escherichia coli physical map. Nucleic Acids Res. **19:**4241–4246.
38. **Karlin, S., and J. Mrázek.** 1996. What drives codon usage in human genes? J. Mol. Biol. **262:**459–472.
39. **Karlin, S., and J. Mrázek.** 1997. Prokaryotic genome-wide comparisons and evolutionary implications. *In* F. J. de Bruijn, J. Lupski, and G. Weinstock (ed.), Bacterial genomes: physical structure and analysis, in press. Chapman & Hall, New York, N.Y.
40. **Karlin, S., and J. Mrázek.** 1997. Compositional biases in eukaryotic genomes. *In* R. S. Verma (ed.), Advances in genome biology, vol. V. Genes and genome, in press. JAI Press Inc., Greenwich, Conn.
41. **Karlin, S., J. Mrázek, and A. M. Campbell.** 1996. Frequent oligonucleotides and peptides of the *Haemophilus influenzae* genome. Nucleic Acids Res. **24:**4263–4272.
42. **Kunkel, T. A.** 1992. Biological asymmetries and the fidelity of eukaryotic DNA replication. Bioessays **14:**303–308.
43. **Lake, J. A.** 1989. Origin of the eukaryotic nucleus: eukaryotes and eocytes are genotypically related. Can. J. Microbiol. **35:**109–118.
44. **Lake, J. A.** 1994. Reconstructing evolutionary trees from DNA and protein sequences: paralinear distances. Proc. Natl. Acad. Sci. USA **91:**1455–1459.
45. **Li, W.-H.** 1996. Molecular evolution. Sinauer Associates, Sunderland, Mass.
46. **Lorenz, M. G., and W. Wackernagel.** 1994. Bacterial gene transfer by natural genetic transformation in the environment. Microbiol. Rev. **58:**563–602.
47. **Maniloff, J.** 1992. Phylogeny of mycoplasmas, p. 549–559. *In* J. Maniloff, R. N. McElhaney, L. R. Finch, and J. B. Baseman (ed.), Mycoplasmas: molecular biology And pathogenesis. American Society for Microbiology, Washington, D.C.
48. **Mongold, J. A.** 1992. DNA repair and the evolution of transformation in *Haemophilus influenzae*. Genetics **132:**893–898.
49. **Moxon, E. R., P. B. Rainey, M. A. Nowak, and R. E. Lenski.** 1994. Adaptive evolution of highly mutable loci in pathogenic bacteria. Curr. Biol. **4:**24–33.
50. **Nei, M.** 1987. Molecular evolutionary genetics. Columbia University Press, New York, N.Y.
51. **Olsen, G. J., C. R. Woese, and R. Overbeek.** 1994. The winds of (evolutionary) change: breathing new life into microbiology. J. Bacteriol. **176:**1–6.
52. **Otwinowski, Z., R. W. Schewitz, G.-G. Zhang, C. L. Lawson, A. Joachimiak, R. Q. Marmorstein, B. F. Luisi, and P. B. Sigler.** 1988. Crystal structure of trp repressor/operator complex at atomic resolution. Nature **335:**321–329.
53. **Phillips, G. J., J. Arnold, and R. Ivarie.** 1987. Mono- through hexanucleotide composition of the *Escherichia coli* genome: a Markov chain analysis. Nucleic Acids Res. **15:**2611–26.
54. **Rafferty, J. B., W. S. Somers, I. Saint-Girons, and S. E. V. Phillips.** 1989. Three-dimensional crystal structures of Escherichia coli met repressor with and without corepressor. Nature **341:**705–710.
55. **Rivera, M. C., and J. A. Lake.** 1992. Evidence that eukaryotes and eocyte prokaryotes are immediate relatives. Science **257:**74–76.
55a. **Roberts, R.** Personal communication.
56. **Roger, A. G., and J. R. Brown.** 1996. A chimeric origin for eukaryotes re-examined. Trends Biochem. Sci. **21:**370–372.
57. **Russell, G. J., and J. H. Subak-Sharpe.** 1977. Similarity of the general designs of protochordates and invertebrates. Nature **266:**533–536.
58. **Russell, G. J., P. M. Walker, R. A. Elton, and J. H. Subak-Sharpe.** 1976. Doublet frequency analysis of fractionated vertebrate nuclear DNA. J. Mol. Biol. **108:**1–23.
59. **Solomon, J. M., and A. D. Grossman.** 1996. Who's competent and when: regulation of natural genetic competence in bacteria. Trends Genet. **12:**150–155.
60. **Tiboni, O., P. Cammarano, and A. M. Sanangelantoni.** 1993. Cloning and sequencing of the gene encoding glutamine synthetase I from the archaeum *Pyrococcus woesei*: anomalous phylogenies inferred from analysis of archaeal and bacterial glutamine synthetase I sequences. J. Bacteriol. **175:**2961–2969.
61. **Travers, A. A.** 1993. DNA-protein interactions. Chapman & Hall, New York, N.Y.
62. **Viale, A. M., and A. K. Arakaki.** 1994. The chaperone connection to the origins of the eukaryotic organelles. FEBS Lett. **341:**146–151.
63. **Woese, C. R.** 1987. Bacterial evolution. Microbiol. Rev. **5:**221–271.
64. **Woese, C. R.** 1996. Phylogenetic trees: whither microbiology. Curr. Biol. **6:**1060–1062.
65. **Woese, C. R., O. Kandler, and M. L. Wheelis.** 1990. Towards a natural system of organisms: proposal for the domains Archaea, Bacteria, and Eucarya. Proc. Natl. Acad. Sci. USA **87:**4576–4579.
66. **Wolffe, A.** 1992. Chromatin structure and function. Academic Press, San Diego, Calif.