J Mol Evol (1998) 47:691–696

JOURNAL OF MOLECULAR EVOLUTION

Base Composition Skews, Replication Orientation, and Gene Orientation in 12 Prokaryote Genomes

Michael J. McLean, Kenneth H. Wolfe, Kevin M. Devine

Department of Genetics, University of Dublin, Trinity College, Dublin 2, Ireland

Received: 2 February 1998 / Accepted: 15 June 1998

Abstract. Variation in GC content, GC skew and AT skew along genomic regions was examined at third codon positions in completely sequenced prokaryotes. Eight out of nine eubacteria studied show GC and AT skews that change sign at the origin of replication. The leading strand in DNA replication is G-T rich at codon position 3 in six eubacteria, but C-T rich in two *Mycoplasma* species. In *M. genitalium* the AT and GC skews are symmetrical around the origin and terminus of replication, whereas its GC content variation has been shown to have a centre of symmetry elsewhere in the genome. *Borrelia burgdorferi* and *Treponema pallidum* show extraordinary extents of base composition skew correlated with direction of DNA replication. Base composition skews measured at third codon positions probably reflect mutational biases, whereas those measured over all bases in a sequence (or at codon positions 1 and 2) can be strongly affected by protein considerations due to the tendency in some bacteria for genes to be transcribed in the same direction that they are replicated. Consequently in some species the direction of skew for total genomic DNA is opposite to that for codon position 3.

Key words: Base composition — Skews — Replication orientation — Gene orientation — Prokaryote genomes

Several reports have addressed the issue of base composition bias in bacterial genomes. The possible causes of such biases have been reviewed by Francino and Ochman (1997) and Mrázek and Karlin (1998) and probably include asymmetry in biochemical processes such as DNA replication and repair (Sueoka 1962; Muto and Osawa 1987) and mutation of the nontranscribed strand during transcription. Lobry (1996a) showed that significant GC skew [the quantity (G − C)/(G + C), measured around the genome using a sliding window] existed in the genome of *Haemophilus influenzae* and the then-sequenced parts of the *Escherichia coli* and *Bacillus subtilis* genomes. The direction of skew switches at the genome's origin and terminus of replication, such that the leading strand in replication is always richer in G than C. This was subsequently confirmed for the complete *B. subtilis* sequence by Kunst et al. (1997) and for *E. coli* by Blattner et al. (1997). *E. coli* was also reported by Lobry (1996a) to show weak AT skew as well as GC skew. A third parameter, GC content in silent codon positions, has been shown to vary systematically around the genome in *Mycoplasma genitalium* (Kerr et al. 1997; McInerney 1997) but not in *Mycoplasma pneumoniae* (Kerr et al. 1997). A fourth measure, cumulative excess of purine (or keto) bases along the genome, was introduced by Freeman et al. (1998). The different methods used by different groups has, however, made it difficult to compare their results.

We analyzed 12 complete prokaryotic genome sequences (9 eubacteria and 3 archaea; Table 1) to investigate how general such biases are, using consistent methods for each genome to permit comparisons among them. In particular, we smoothed the data by using a larger window size (300,000 nucleotides) than in previous studies and concentrated on the third positions of codons (which are more likely to show mutational influences than first or second positions). When third posi-

Correspondence to: K.M. Devine; *e-mail:* kdevine@tcd.ie

Table 1. Nucleotide composition statistics in 12 prokaryotic genomes[a]

	Eubacteria									Archaebacteria		
	Bacillus subtilis	*Borrelia burgdorferi*	*Trepo-nema pallidum*	*Escher-ichia coli*	*Haemo-philus influenzae*	*Helico-bacter pylori*	*Myco-plasma genitalium*	*Myco-plasma pneumoniae*	*Synecho-cystis* sp.	*Archeo-globus fulgidus*	*Methano-bacterium thermoautotrophicum*	*Methano-coccus jannashii*
Genome length (×000 bp)	4215	911	1138	4639	1830	1668	580	816	3573	2178	1751	1665
Average GC content (all bases), %	44	29	53	51	38	39	32	40	48	49	50	31
Independent origin source												
Confirmed experimentally	Y			Y								
Putative sequence found					Y							
Putative from dnaA homology		Y	Y			Y	Y	Y	Y			
Terminus Source												
Confirmed experimentally	Y			Y								
Inferred		Y	Y		Y	Y	Y	Y	Y			
Fig. 1 GC3 skew quality	Strong	Strong	Strong	Strong	Strong	Strong	Strong	Strong	None	None	Weak	None
Fig. 1 AT3 skew quality	Strong	Strong	Strong	Strong	Weak	Weak	Strong	Strong	None	None	Weak	None
GC skew sign reversal							Y	Y				
AT skew sign reversal	Y						Y	Y				
% genes transcribed in replicn. dirn.	74	65	61	54	54	56	78	77	50	—	—	—
% ribosomal proteins in replicn. dirn.	94	96	92	93	87	88	98	98	83	—	—	—
GC1 skew (leading : lagging) genes, %	30 25	47 33	22 12	19 16	31 27	36 32	29 30	25 23	20 19	40 40	35 31	58 58
GC2 skew (leading : lagging) genes, %	−16 −19	−9 −25	−8 −16	−10 −12	−15 −16	−8 −11	−16 −14	−16 −14	−12 −12	−5 −4	−8 −5	−7 −7
GC3 skew (leading : lagging) genes, %	9 −5	33 −29	25 −8	8 −1	8 −6	1 −10	−6 −8	−9 −11	−6 −5	4 5	−5 −9	25 22
AT1 skew (leading : lagging) genes, %	27 24	21 37	7 15	23 24	15 18	23 23	28 28	23 24	17 16	34 34	37 37	29 29
AT2 skew (leading : lagging) genes, %	4 0	1 9	−7 −4	−3 −4	1 3	7 6	7 5	8 7	−1 −1	−2 −3	−5 −5	5 6
AT3 skew (leading : lagging) genes, %	−4 −3	−23 19	−27 −8	−20 −17	−9 −5	−10 −10	−12 −9	−12 −9	−14 −14	−6 −5	7 4	6 6

[a] Sequence data (Fleischmann et al. 1995; Fraser et al. 1995, 1997; Bult et al. 1996; Himmelreich et al. 1996; Kaneko et al. 1996; Blattner et al. 1997; Klenk et al. 1997; Kunst et al. 1997; Smith et al. 1997; Tomb et al. 1997) were obtained from the TIGR WWW site (http://www.tigr.org) and links therefrom. Because several of the sequences were completely unannotated at the time of analysis (the *Borrelia burgdorferi* and *Treponema pallidum* sequences were provisional and may contain minor errors), and to maintain consistency of approach to all genomes, we used a single simple ORF-finding program to generate a crude annotation of all 12 genomes. ORFs beginning with ATG and GTG start codons, and larger than 100 codons, were included provided that they did not overlap by >50 bp with a larger ORF. Comparison of these annotations to author-supplied annotations, where available, revealed only minor differences which will not affect the results. Origins and termini of replication have been determined experimentally for *E. coli* and *B. subtilis*. Origins for other species were inferred from the location of a *dnaA* gene. Termini locations in eubacteria were inferred from a change in skew sign (Lobry 1996b) except for *Synechocystis*, where it was assumed to be 180° from the origin. The ''quality'' of skews was assessed subjectively from the plots in Fig. 1. Ribosomal protein genes were identified by BLASTP searches using as queries the 56 *E. coli* ribosomal proteins named in Swissprot. The bottom six lines show GC skew and AT skew at codon positions 1, 2, and 3; two numbers, indicating separately the skew in leading and lagging replication strand genes, are given for each organism. For example, GC skew at codon position 3 in *Borrelia burgdorferi* is +33 in leading-strand genes and −29 in lagging-strand genes.

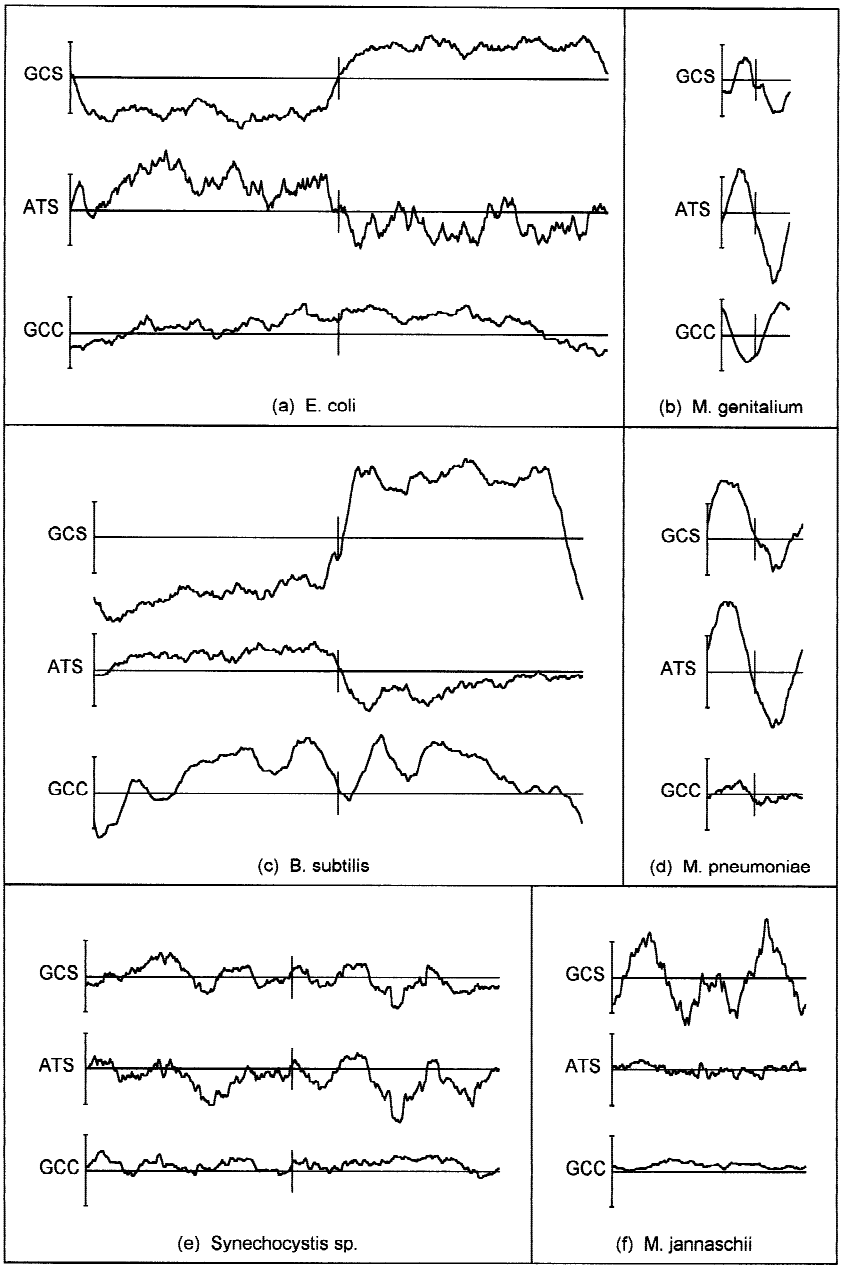

Fig. 1. Third codon position GC skew (GCS), AT skew (ATS), and GC content (GCC) in 12 completely sequenced prokaryotic genomes. All plots are at the same scale. The vertical bar at the *left* of each plot extends 5% above and below the axis. The window size was 300,000 nucleotide sites and the step size was 10,000 sites. Eubacteria are plotted with the putative origin of replication (marked by a *vertical bar*) at the center, and so the sequence may be rotated compared to the original published numbering scheme.

tions of codons are examined, there is strong GC skew in eight of the nine eubacteria, and AT skew is strong in six eubacteria and weak in two others (Fig. 1). Only *Synechocystis* and the three archaea do not show strong skews. In all cases where strong skews at third codon positions exist, they switch sign at the probable origin and terminus of replication. The leading strand of replication is comparatively G–T rich in all eubacteria (see also Perrière et al. 1996; Francino and Ochman 1997) except the two *Mycoplasma* species, where it is C–T rich. In *M. genitalium* (Fig. 1b) the GC and AT skews change sign at the origin and terminus of replication (Lobry 1996b), whereas the silent codon position G + C content has maximum and minimum values at apparently featureless but diametrically opposing points in the genome (Kerr et al. 1997; McInerney 1997).

By far the largest skews (up to 30%) were exhibited by the two spirochaetes, *Borrelia burgdorferi* (Fraser et al. 1997) and *Treponema pallidum.* Since these 2 organisms also exhibit the lowest and highest GC contents, respectively, of the 12 organisms analyzed (Table 1), it appears that the pressure that creates skew is largely independent of the pressure that determines GC content. Remarkably, overall third-position G + C content in the spirochaetes is almost invariant along their genomes despite the uneven distribution of G and C between the two DNA strands. The two spirochaetes have similar skew patterns despite having different chromosome structures

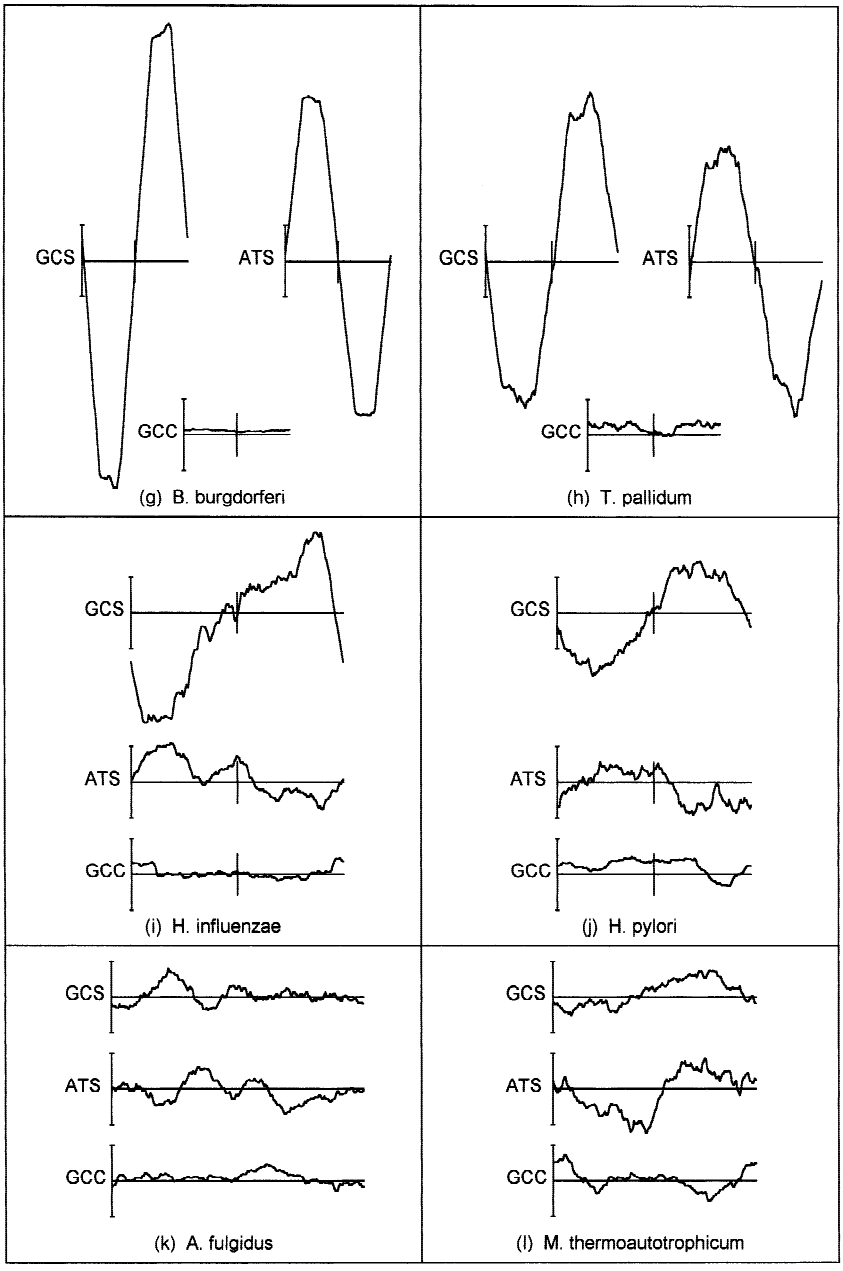

Fig. 1. Continued.

(linear versus circular) and little conservation of gene order.

The last six rows in Table 1 show GC skews and AT skews for each codon position, with genes on the leading and lagging strands shown separately. In all 12 organisms the first codon positions show strong positive GC and AT skews, while the second positions show weaker negative GC skews and mixed-sign but very much weaker AT skews. When only first and second codon positions are considered, the corresponding skews in genes encoded on leading and lagging strands are of the same sign and approximately the same magnitude. The differences between leading and lagging strands reflect the combined effects of replication- and transcription-induced mutation (Lobry 1996a; Francino and Ochman 1997), while the small magnitude of these differences reflects selective pressure to preserve the same amino acids. Mutations induced by replication will increase the skew of genes encoded on one strand and decrease the skew of genes on the other strand (hence the differences in the two values). Mutations induced by transcription generally result in the skews of genes on the leading and lagging strands either both increasing or both decreasing (so there is no change in the difference observed), provided that the transcriptional orientation of genes on the chromosome is random. However, if a disproportionate number of heavily expressed genes are coded on the leading strand (Brewer 1988; Blattner et al. 1997) (Table 1), the leading-strand genes may sustain more transcription-induced skew than the lagging-strand genes, which

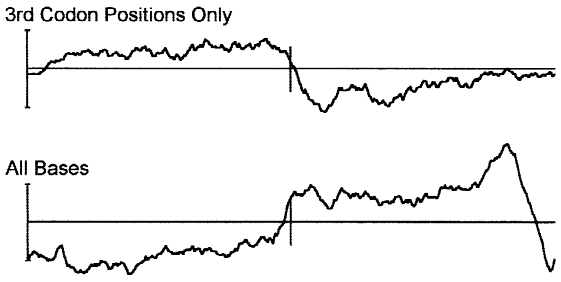

Fig. 2. AT skew in *Bacillus subtilis* calculated for third codon positions only (same as Fig. 1c) and for all bases in the genome.

should be observable. Unfortunately, there is insufficient information to distinguish the amount of skew induced by replication from that induced by transcription. At third codon positions the selective pressure is greatly decreased, with the result that the differences caused by mutation are greatly increased, even to the extent that the signs of the skews on opposite strands are often different.

The asymmetry between the two DNA strands is sufficient in some cases to cause significant differences in the distributions of both codons and amino acids in their genes. For example, we observed that in *B. burgdorferi*, the amount of each codon (as a fraction of total codons) differs by an average of 40% between the genes encoded on the two strands (results not shown), and the corresponding difference for amino acids is 19%. We speculate that this observation will provide a means of improving the effectiveness of gene prediction programs such as GLIMMER (Salzberg et al. 1998), which calculate a statistical model based on the characteristics of known genes and use the result to recognize and predict other genes. We envisage programs that calculate and use separate statistical models for leading- and lagging-strand genes.

Finally, we note that strong base composition skews in ''total'' genomic DNA (rather than third codon positions) can arise due to nonrandomness in the transcriptional orientation of genes on the chromosome, even in the absence of skews induced by replication or transcription. Blattner et al. (1997) reported that in *E. coli*, GC skew patterns similar to that seen at codon position 3 (Fig. 1a) are also found in the total sequence (and at codons positions 1 and 2). When we investigated this for other species we found that in some cases the direction of the skew for total DNA was opposite to the direction for codon position 3. This occurs in the two *Mycoplasma* species for GC and AT skew and in *B. subtilis* for AT skew (Fig. 2). These are also the species with the strongest tendency to arrange genes such that their transcriptional orientation is the same as their replication direction (Table 1). This tendency becomes even more marked if only highly expressed genes such as ribosomal protein genes are considered (Table 1) (Brewer 1988). The requirement that genes encode proteins causes biases in the base composition of codon positions 1 and 2 on the sense strand of genes, even in the absence of any skews induced by replication and/or transcription. The skews presented in Figs. 1 and 2 combine the skews of genes on the leading and lagging strands. If genes are oriented randomly on the chromosome, the leading strand in replication will contain approximately equal numbers of sense and antisense strands of genes, and so the influence of amino acid composition on genomic base composition will cancel itself out. However, if genes are not oriented randomly, the base composition of the leading and lagging strands will be affected by amino acid composition considerations, causing skews (such as the sign switch shown in Fig. 2) that have nothing to do with mutational biases. This point was also made very recently by Mrázek and Karlin (1998).

We believe that this combination of constraint on amino acid sequences and nonrandom gene orientation is also the principal cause of the correlation, reported recently by Freeman et al. (1998) for many bacterial genomes, between the cumulative excess of purine bases (measured in total DNA) and the cumulative excess of bases on the coding strand. Freeman et al. proposed that the correlation they found resulted from ''asymmetrical errors in DNA synthesis,'' but we note here that an excess of purines on the coding strand is a universal feature of genes from all organisms, including those with multiple replication origins and seemingly random gene orientations (human, mouse, yeast, *Arabidopsis*, *Dictyostelium*) as well as all 12 prokaryotes studied here. Coding strands in large sets of genes from all these organisms have average purine contents between 51 and 59%, which must reflect amino acid constraints rather than mutational biases.

Acknowledgment. We thank J. Lobry for helpful comments and suggestions.

References

Blattner FR, Plunkett G, Bloch CA, Perna NT, Burland V, Riley M, Collado-Vides J, Glasner JD, Rode CK, Mayhew GF, et al. (1997) The complete genome sequence of *Escherichia coli* K-12. Science 277:1453–1474

Brewer BJ (1988) When polymerases collide: Replication and the transcriptional organization of the *E. coli* chromosome. Cell 53:679–686

Bult CJ, White O, Olsen GJ, Zhou L, Fleischmann RD, Sutton GG, Blake JA, FitzGerald LM, Clayton RA, Gocayne JD, et al. (1996) Complete genome sequence of the methanogenic archaeon, *Methanococcus jannaschii.* Science 273:1058–1073

Fleischmann RD, Adams MD, White O, Clayton RA, Kirkness EF, Kerlavage AR, Bult CJ, Tomb JF, Dougherty BA, Merrick JM, et al. (1995) Whole-genome random sequencing and assembly of *Haemophilus influenzae Rd.* Science 269:496–512

Francino MP, Ochman H (1997) Strand asymmetries in DNA evolution. Trends Genet 13:240–245

Fraser CM, Gocayne JD, White O, Adams MD, Clayton RA, Fleischmann RD, Bult CJ, Kerlavage AR, Sutton G, Kelley JM, et al.

(1995) The minimal gene complement of *Mycoplasma genitalium.* Science 270:397–403

Fraser CM, Casjens S, Huang WM, Sutton GG, Clayton R, Lathigra R, White O, Ketchum KA, Dodson R, Hickey EK, et al. (1997) Genomic sequence of a Lyme disease spirochaete, *Borrelia burgdorferi.* Nature 390:580–586

Freeman JM, Plasterer TN, Smith TF, Mohr SC (1998) Patterns of genome organization in bacteria. Science 279:1827a

Himmelreich R, Hilbert H, Plagens H, Pirkl E, Li BC, Herrmann R (1996) Complete sequence analysis of the genome of the bacterium *Mycoplasma pneumoniae.* Nucleic Acids Res 24:4420–4449

Kaneko T, Sato S, Kotani H, Tanaka A, Asamizu E, Nakamura Y, Miyajima N, Hirosawa M, Sugiura M, Sasamoto S, et al. (1996) Sequence analysis of the genome of the unicellular cyanobacterium *Synechocystis* sp. strain PCC6803. II. Sequence determination of the entire genome and assignment of potential protein-coding regions. DNA Res 3:185–209

Kerr AR, Peden JF, Sharp PM (1997) Systematic base composition variation around the genome of *Mycoplasma genitalium*, but not *Mycoplasma pneumoniae.* Mol Microbiol 25:1177–1179

Klenk HP, Clayton RA, Tomb JF, White O, Nelson KE, Ketchum KA, Dodson RJ, Gwinn M, Hickey EK, Peterson JD, et al. (1997) The complete genome sequence of the hyperthermophilic, sulphate-reducing archaeon *Archaeoglobus fulgidus.* Nature 390:364–370

Kunst F, Ogasawara N, Moszer I, Albertini AM, Alloni G, Azevedo V, Bertero MG, Bessieres P, Bolotin A, Borchert S, et al. (1997) The complete genome sequence of the gram-positive bacterium *Bacillus subtilis.* Nature 390:249–256

Lobry JR (1996a) Asymmetric substitution patterns in the two DNA strands of bacteria. Mol Biol Evol 13:660–665

Lobry JR (1996b) Origin of replication of *Mycoplasma genitalium.* Science 272:745–746

McInerney JO (1997) Prokaryotic genome evolution as assessed by multivariate analysis of codon usage patterns. Microbial Comp Genom 2:1–10

Mrázek J, Karlin S (1998) Strand compositional asymmetry in bacterial and large viral genomes. Proc Natl Acad Sci USA 95:3720–3725

Muto A, Osawa S (1987) The guanine and cytosine content of genomic DNA and bacterial evolution. Proc Natl Acad Sci USA 84:166–169

Perrière G, Lobry JR, Thioulouse J (1996) Correspondence discriminant analysis: A multivariate method for comparing classes of protein and nucleic acid sequences. Comput Appl Biosci 12:519–524

Salzberg SL, Delcher AL, Kasif S, White O (1998) Microbial gene identification using interpolated Markov models. Nucleic Acids Res 26:544–548

Smith DR, Doucette-Stamm LA, Deloughery C, Lee H, Dubois J, Aldredge T, Bashirzadeh R, Blakely D, Cook R, Gilbert K, et al. (1997) Complete genome sequence of *Methanobacterium thermoautotrophicum* deltaH: Functional analysis and comparative genomics. J Bacteriol 179:7135–7155

Sueoka N (1962) On the genetic basis of variation and heterogeneity of DNA base composition. Proc Natl Acad Sci USA 48:1141–1149

Tomb JF, White O, Kerlavage AR, Clayton RA, Sutton GG, Fleischmann RD, Ketchum KA, Klenk HP, Gill S, Dougherty BA, et al. (1997) The complete genome sequence of the gastric pathogen *Helicobacter pylori.* Nature 388:539–547