Molecular Microbiology (1999) **32**(1), 11–16

MicroGenomics

Universal replication biases in bacteria

Eduardo P. C. Rocha,[1,2] Antoine Danchin[2] and Alain Viari[1,3]*

[1]*Atelier de BioInformatique, Université Paris VI, 12, rue Cuvier 75005 Paris, France.*

[2]*Régulation de l'Expression Génétique, Institut Pasteur, 28 rue du Dr Roux, 75724 Paris, France.*

[3]*Physico-Chimie-Curie, Institut Curie, 26 rue d'Ulm, 75005 Paris, France.*

Summary

Analysis of 15 complete bacterial chromosomes revealed important biases in gene organization. Strong compositional asymmetries between the genes lying on the leading versus lagging strands were observed at the level of nucleotides, codons and, surprisingly, amino acids. For some species, the bias is so high that the sole knowledge of a protein sequence allows one to predict with almost no errors whether the gene is transcribed from one strand or the other. Furthermore, we show that these biases are not species specific but appear to be universal. These findings may have important consequences in our understanding of fundamental biological processes in bacteria, such as replication fidelity, codon usage in genes and even amino acid usage in proteins.

Introduction

Bacterial chromosome replication usually starts at a single origin, and two replication forks propagate in opposite directions up to termination signals (Marians, 1992). As the replication mechanism differs for the two strands of the duplex DNA (Marians, 1992), this process may, in principle, give rise to compositional asymmetries between the leading (continuously replicated) strand and the lagging strand. Indeed, numerous indications of these asymmetries have been shown recently, essentially reflecting an excess of G over C in the leading strand (Lobry, 1996; Francino and Ochman, 1997; Freeman *et al*., 1998; Mrázek and Karlin, 1998). Most of these studies focus on the nucleotide genome composition, using variants of the (G−C)/(G+C) plots initially proposed by Lobry (1996), although recent analyses using oligonucleotides have been reported (Salzberg *et al*., 1998). In contrast, few studies have been undertaken to investigate the impact of this bias at the level of genes (Lobry, 1996; Kerr *et al*., 1997; McInerney, 1998), and none have been undertaken at the level of proteins. We have therefore undertaken a study to address the following questions: (i) are there detectable compositional asymmetries in genes in terms of nucleotides, codons or amino acids? (ii) how are they related to the known G/C bias? (iii) are they species specific or universal? The main difficulty in answering these questions is technical: when considering more than the four variables G, A, T, C (for instance, considering the abundance of codons in genes), it becomes virtually impossible to check all possible combinations of variables manually, and more sophisticated techniques are required. For this purpose, we have introduced the use of a classical statistical tool called linear discriminant analysis (LDA), initially formulated by R. A. Fisher in 1936 (Fisher, 1936). Given two populations of objects described by a set of n variables x_i, the goal is to build a function $F(x) = \alpha_0 + \sum_{i=1}^{n} \alpha_i x_i$, such that $F(x) > 0$ when x belongs to the first population and $F(x) < 0$ when x belongs to the second population. The purpose of LDA is to determine the coefficients $\{\alpha_i\}_{i=1,n}$ that discriminate 'best' between the two populations [in the context of LDA, 'best' means optimized ratio (mean difference)^2/variance]. The application of this technique to our problem is straightforward. Let us set a putative origin of replication at an arbitrary position p in the chromosome (Fig. 1). This creates two populations of genes: those lying on the leading strand and those lying on the lagging strand. Then, we describe each gene by using a set of variables (for instance, the relative frequencies of the 61 non-stop codons), and we subject the two populations of genes to LDA in order to evaluate how accurately they can be separated (i.e. predicted) on the basis of these variables (Fig. 1). By varying the position p along the chromosome, we plot the variation of the prediction accuracy. If any bias does exist, we expect higher accuracy when p coincides with the origin or termination region, as all genes 'add up' their effect at these locations. Here, we report the results obtained using four different sets of variables: (i) the relative frequency of nucleotides (four variables); this particular analysis is equivalent to, although mathematically different from, the

Received 31 December, 1998; accepted 6 January, 1999. *For correspondence. E-mail Alain.Viari@snv.jussieu.fr; Tel. (+33) 1 44 27 65 36; Fax (+33) 1 44 27 63 12.

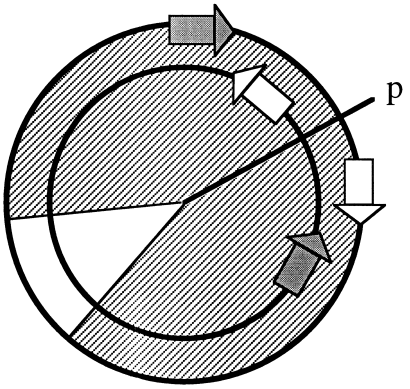

Fig. 1. Schema of the analysis. A putative origin of replication is set at an arbitrary position p in the chromosome, and we consider the genes lying on the leading strand (white arrows) and on the lagging strand (grey arrows) within a window centred on p and covering 7/8th of the chromosome. The reason for this is that, unlike the origin, termination may occur at several Ter sites opposite to the origin (Marians, 1992). This window therefore allows for some uncertainty in the termination loci. For a linear chromosome (the only case studied here is *B. burgdorferi*), this is not required, and the window covers the whole chromosome. Within each of the two populations of genes, we draw 70% of the elements to be subjected to linear discriminant analysis randomly (learning set). The remaining 30% (test set) are used to evaluate the quality of the discrimination, as measured by the 'accuracy', which is simply the percentage of correct predictions on the test set (more sophisticated measures, which take into account a possible imbalance in the size of the populations, have also been tested and do not significantly change the results presented here).

above-mentioned studies (Lobry, 1996; Grigoriev, 1998); (ii) the relative frequency of nucleotides at each position of the codon (12 variables); (iii) the relative frequency of codons (61 variables); and (iv) the relative frequency of amino acids (20 variables). Using these four sets of variables, we investigated the 15 complete and annotated prokaryotic chromosomes published to date, namely three archeae, *Archaeoglobus fulgidus* (Klenk *et al*., 1997), *Methanococcus jannaschii* (Bult *et al*., 1996) and *Methanobacterium thermoautotrophicum* (Smith *et al*., 1997) and 12 eubacteria, *Aquifex aeolicus* (Deckert *et al*., 1998), *Bacillus subtilis* (Kunst *et al*., 1997), *Borrelia burgdorferi* (Fraser *et al*., 1997), *Chlamydia trachomatis* (Stephens *et al*., 1998), *Escherichia coli* (Blattner *et al*., 1997), *Haemophilus influenzae* (Fleischmann *et al*., 1995), *Helicobacter pylori* (Tomb *et al*., 1997), *Mycobacterium tuberculosis* (Cole *et al*., 1998), *Mycoplasma genitalium* (Fraser *et al*., 1995), *Mycoplasma pneumoniae* (Himmelreich *et al*., 1996), *Synechocystis* sp. (Kaneko *et al*., 1996) and *Treponema pallidum* (Fraser *et al*., 1998).

Results and discussion

For nine out of the 15 species, the plots exhibit two clear maxima located at or near the origin and terminus of replication (Fig. 2). This clearly demonstrates the existence of a bias acting at the level of nucleotides, codons and, surprisingly, also amino acids. The amplitude of this bias was totally unexpected and is very strong, judging from the absolute value of the prediction accuracy (*y*-axis of the plots). For instance, codon bias generally gives 70–80% correct predictions (Fig. 2), whereas LDA random prediction would yield only 50% accuracy. Extreme cases are obtained for the genomes of *B. burgdorferi*, *T. pallidum* and *C. trachomatis*. In the case of *B. burgdorferi*, the accuracy of discrimination using the amino acid frequencies is 96%. In other words, the sole knowledge of a protein sequence of this species is sufficient to predict with almost no errors (4% of false predictions) the orientation of the corresponding gene with respect to replication. *B. burgdorferi* is an extreme case, but it is clear that such a bias also exists for the other species (Fig. 2). One should note that a strong bias in the codon usage of *B. burgdorferi* was revealed recently by McInerney (1998) using factorial analysis, but its implications have not yet been perceived at the protein level. The six other species reveal ambiguous plots (*M. jannaschii*, *M. genitalium* and *M. pneumoniae*) or no significant bias at all (*A. aeolicus*, *A. fulgidus* and *Synechocystis* sp.). It is interesting to note that no experimentally determined origin of replication has yet been proposed for the last three cases. It should be pointed out that the absence of a significant skew is not a distinctive feature of archeae, as *M. thermoautotrophicum* displays a clear codon bias (Fig. 2). The absence of observable asymmetries may suggest that, like eukaryotes, replication could take place at different locations on the chromosome of these species or, alternatively, that these genomes are shuffled so frequently that biases do not have time to become established. To elucidate the nature of the bias observed in the nine species, we need to examine more closely the coefficients of the discriminant analysis when p is set to the known origin of replication (or to the maximum of the previous curves when experimentally unknown) (Table 1). Indeed, positive coefficients indicate variables in favour of the leading strand and, conversely, negative coefficients correspond to variables in favour of the lagging strand. Moreover, the amplitude of the coefficients indicates the importance of the corresponding variable in the discrimination. Examination of the first set of variables (nucleotides) confirms earlier results on the excess of G over C in the leading strand (Table 1). Moreover, closer examination reveals that T is constantly dominant over A on the leading strand. This answers a still open question about the status of A and T (Mrázek and Karlin, 1998), showing (i) that A and T are also skewed and (ii) that the bias is keto (GT) versus amino (AC) rather than purine versus pyrimidine. The second set of variables gives a more focused view of the bias within genes: there is a clear universal skew in favour of G over C in the third position to the codon (Table 1). This observation is consistent with the fact that, in the genetic

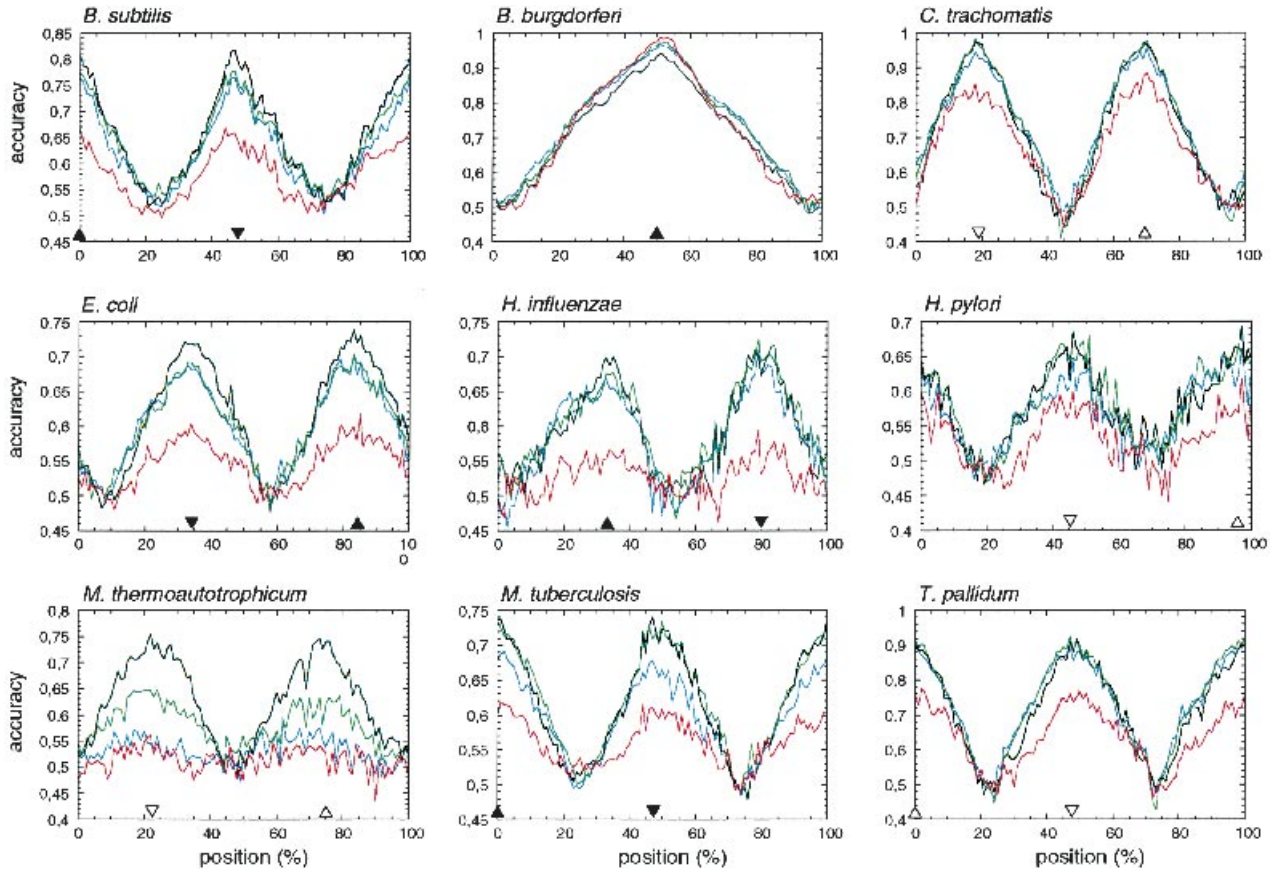

Fig. 2. Plots of the prediction accuracy (*y*-axis) when varying the location of the origin of replication (*x*-axis). The position on the *x*-axis is expressed as the percentage of total chromosome length. For each species, we plot the accuracy obtained using four sets of variables to describe genes: blue, relative frequency of the four nucleotides; green, relative frequency of nucleotides at each position within codons; black, relative frequency of codons; and red, relative frequency of amino acids in the encoded proteins. Black triangles on the *x*-axis represent known origins (pointing upwards) and termini (pointing downwards) of replication. White triangles represent origins and termini proposed in this study.

code, the third position is more tolerant to mutations (Muto and Osawa, 1987). However, a closer examination shows that the keto-amino bias sometimes extends up to the first base and, more rarely, to the second base of the codon (Table 1). Indeed, the examination of the third set of variables (codons) shows that the previous rule does not apply uniformly on all codons (Table 1). The leading strand is clearly biased towards codons starting or ending with G or T but (i) not all such codons are selected for and (ii) some codons are chosen universally. The most conspicuous codons are GTG (V), GCG (A), GAG (E) on the leading strand and CTC (L), GCC (A), CCC (P), ATC (I) and ACC (T) on the lagging strand (Table 1). A comparison of the most discriminative codons with the overall codon usage of each species did not reveal any correlation. This eliminates the possibility that the bias is caused by the well-known over-representation of highly expressed genes on the leading strand (Brewer, 1988; Francino and Ochman, 1997). Therefore, the codon asymmetry revealed here is distinct from the usual codon bias resulting from gene expression levels. Finally, the most unexpected finding is that there is a constant trend towards valine on the leading strand and, to a lesser extent, towards threonine and isoleucine on the lagging strand (Table 1). As far as we know, this is the first time that a constraint acting at the DNA level (replication) has been shown to have such a deep impact on protein composition. This impact can be revealed more dramatically in the extreme cases of *B. burgdorferi*, *C. trachomatis* and *T. pallidum* by simply plotting the abundance of valine versus threonine in the translation product of each gene (Fig. 3). These plots reveal that the constraint is so high that it enforces a complete inversion of the valine–threonine ratio in the two strands. An interesting illustration of this last fact is provided by two homologous genes (BB0629 and BB0408) in *B. burgdorferi* that are highly similar to the *E. coli fruA* gene encoding the fructose permease IIBC component. The two corresponding proteins exhibit a high level of identity (38% identities for a total length of 625 amino acids), but BB0408 is located on the leading strand, whereas BB0629 is on the lagging

Table 1. Most discriminating factors between the leading and lagging strand.

	Bases		Codon bases		Codons		Amino acids	
	+	−	+	−	+	−	+	−
B. subtilis	**G**	**C**	**G3**	**C3**	GAG GCG GTG GGG	GCC CCC ACC TTC	**V** E R D	F S L I H **T**
			G1		ACG AAG CCG GAT	CTC TCC ATC		
					GTT GGT GAA CGT			
					GTA			
B. burgdorferi	**G**	A	T3	A3	GTT AAG GAT GGT	ATA AAC CTA ACA	**V** D R	I **T** K N
	T	**C**	**G3**	**C3**	TCT GAG AGT TTG	AAA ATC TAC GAA		
				A1	AGG	CTC CAC TTC GAC		
C. trachomatis	**G**	**C**	**G3**	**C3**	GAG TTG GTT GTG	CTC ACC CTA TCC	**V** R G	**T** P I L S N
			G1	C1	GGG AAG CGT AGT	ATC CTT CGC CAA		
			G2	C2	CGG GGT CAG GAT	TTC AAC CAC		
					AGG GTA			
E. coli	**G**	**C**	**G3**	**C3**	GCG CGT GTG GGG	GCC CCC CTC CTA	**V** G	**T** N I H L P
			G1					
			T3					
H. influenzae	**G**	**C**	T3	**C3**	GTT GCT CGT GTG	ACC GCC CTC CGC	**V**	**T** P N
	T		**G3**		GCG CAG GAG GGT	CCC GTC CTA AAC		
			G1		TCT TTG AAG			
H. pylori	**G**	**C**	**G3**	**C3**	GAG AAG GTG TTG	CCC CTC TAC ACC	**V** K E R S	P **T** F
				C1	AGG GCG	TTC GTC ATC CTT		
						GCC		
M. thermoautotrophicum	**G**	**C**	G1	**C3**	GCA GTA ACG CCA	GCC CGG AGT ACC	**V** A	L Y H Q K W
			C2	C1	CCG AGG GAG GCG	CCC GGT ATC TTG		
			G3		TCA	AAA TAC		
			A3					
M. tuberculosis	**G**	**C**	**G3**	C3	GTG TTG GGT GCG	ACC CCA GCA CCC	**V**	**T**
	T		T3	**A3**	GTT GAT CGG GGG	CAA GCC ACA GGC		
					CGT GAG	CTC CGC AGC CTA		
						TAC AAC ATC TCA		
						CAC CGA		
T. pallidum	**G**	**C**	**G3**	C3	GTG TTG CGT GAT	ACC CTC CAC ACA	**V**	**T** H
	T		T3	**A3**	GTT GCG GGG GGT	GAC GCC AAC TCC		
			G1		CGG	CTA TTC CCC TAC		
						ATC AGC		

For each species (rows), the table indicates the most discriminating variables in each of the four sets. In each table entry, the variables are listed according to decreasing importance, as measured by the amplitude of the corresponding coefficient in the discriminant function (see text). The listing stops when the coefficients drop to less than one half of the highest one. Columns marked '+' contain the factors selected for in the leading strand, and columns marked '−' contain the factors selected for in the lagging strand. We indicate in bold the factors that are found universally across the species (except for codons, for the sake of clarity).

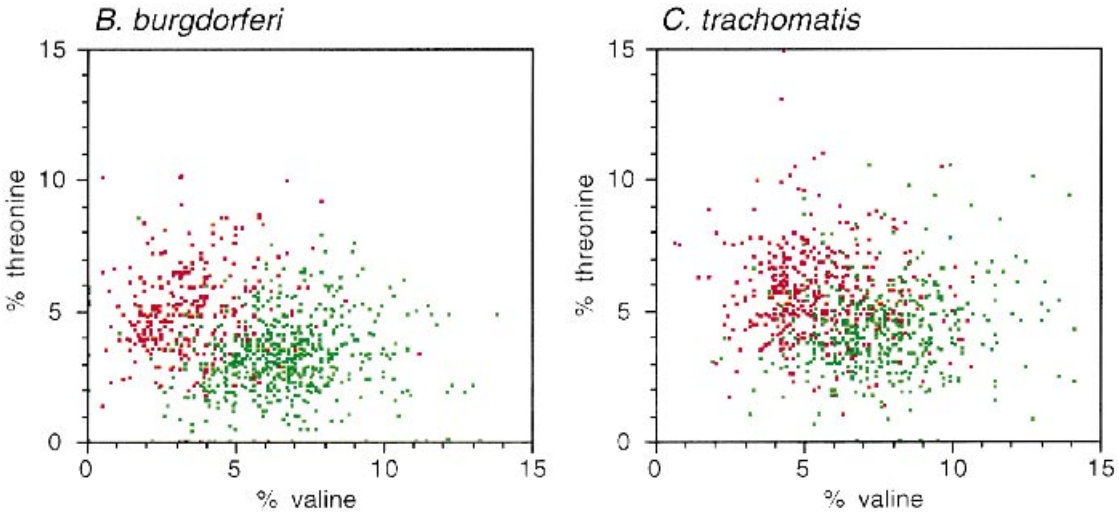

Fig. 3. Evidence of a strong amino acid compositional bias in proteins of *B. burgdorferi* and *C. trachomatis*. The percentage of valine (*x*-axis) is plotted versus threonine (*y*-axis) for all proteins of these two species. Green dots correspond to proteins encoded by genes on the leading strand and red dots to proteins encoded by genes on the lagging strand. A similar plot is obtained using valine versus isoleucine in *B. burgdorferi*.

strand. Examination of the alignment of the two proteins reveals a very strong mutational polarization, as 19 valine to isoleucine mutations are observed in the leading to lagging direction compared with only four valine to isoleucine transitions in the opposite direction.

Besides their fundamental implications for our understanding of the replication process, these results clearly have some practical consequences in the analysis of complete genomes. When looking for putative coding regions using Markov models (Borodovski *et al*., 1994), it is crucial for some species to use a different model for each strand. At the protein level, these results dramatically highlight the limitations of the use of a single, symmetrical scoring matrix such as PAM (Dayhoff *et al*., 1978). For species such as *B. burgdorferi* or *C. trachomatis* for instance, ignoring the facts presented here would almost certainly lead to very misleading conclusions in phylogenetic studies.

Acknowledgements

We thank J. F. Tomb, G. Fichant, E. Coissac and A. Simionovici for their scientific help and friendly support. E.R. acknowledges a fellowship from PRAXIS XXI.

Note added in proof

Recently released complete genome of the bacterium *Rickettsia prowazekii* (Andersson *et al. Nature* **396:** 133–143) reveals exactly the same biases as those described above. With the origin of replication located at 0 kb (as suggested in the original publication), the prediction accuracy, based on the codon bias, is 87%.

References

Blattner, F.R., Plunkett, III, G., Bloch, C.A., Perna, N.T., Burland, V., *et al*. (1997) The complete genome sequence of *Escherichia coli* K-12. *Science* **277:** 1453–1461.

Borodovski, M., Rudd, K.E., and Koonin, E.V. (1994) Intrinsic and extrinsic approaches for detecting genes in a bacterial genome. *Nucleic Acids Res* **22:** 4756–4767.

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

Bult, C.J., White, O., Olsen, G.J., Zhou, L., Fleischmann, R.D., *et al*. (1996) Complete genome sequence of the methanogenic archaeon, *Methanococcus jannaschii*. *Science* **273:** 1058–1072.

Cole, S.T., Brosch, R., Parkhill, J., Garnier, T., Churcher, C., *et al*. (1998) Deciphering the biology of *Mycobacterium tuberculosis* from the complete genome sequence. *Nature* **393:** 537–544.

Dayhoff, M.O., Schwartz, R.M., and Orcutt, B.C. (1978) A model of evolutionary change in proteins. In *Atlas of Protein Sequence and Structure, Vol. 5*. Dayhoff, M.O. (ed.). Washington, DC: National Biomedical Research Foundation, pp. 345–352.

Deckert, G., Warren, P.V., Gaasterland, T., Young, W.G., Lenox, A.L., *et al*. (1998) The complete genome of the hyperthermophilic bacterium *Aquifex aeolicus*. *Nature* **26:** 353–358.

Fisher, R.A. (1936) The use of multiple measurements in taxonomic problems. *Ann Eugen* **7:** 179–188.

Fleischmann, R.D., Adams, M.D., White, O., Clayton, R.A., Kirkness, E.F., *et al*. (1995) Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd. *Science* **269:** 496–512.

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

Fraser, C.M., Casjens, S., Huang, W.M., Sutton, G.G., Clayton, R., *et al*. (1997) Genomic sequence of a Lyme disease spirochaete, *Borrelia burgdorferi*. *Nature* **390:** 580–586.

Fraser, C.M., Gocayne, J.D., White, O., Adams, M.D., Clayton, R.A., *et al*. (1995) The minimal gene complement of *Mycoplasma genitalium*. *Science* **270:** 397–403.

Fraser, C.M., Norris, S.J., Weinstock, G.M., White, O., Sutton, G.G., *et al*. (1998) Complete genome sequence of *Treponema pallidum* the syphilis spirochete. *Science* **281:** 375–388.

Freeman, J.M., Plasterer, T.N., Smith, T.F., and Mohr, S.C. (1998) Patterns of genome organization in bacteria. *Science* **279:** 1827a.

Grigoriev, A. (1998) Analyzing genomes with cumulative skew diagrams. *Nucleic Acids Res* **26:** 2286–2290.

Himmelreich, R., Hilbert, H., Plagens, H., Pirkl, E., Li, B.C., and 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., *et al*. (1996) Sequence analysis of the genome of the unicellular cyanobacterium *Synechocystis* sp strain PCC6803. *DNA Res* **3:** 109–136.

Kerr, A.R.W., Peden, J.F., and Sharp, P.M. (1997) Systematic base composition variation around the genome of *Mycoplasma genitalium*, but not *Mycoplasma pneumoniae*. *Mol Microbiol* **25:** 1177.

Klenk, H.-P., Clayton, R.A., Tomb, J.F., White, O., Nelson, K.E., *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, A.M., Alloni, G., *et al*. (1997) The complete genome sequence of the Gram-positive bacterium *Bacillus subtilis*. *Nature* **390:** 249–256.

Lobry, J.R. (1996) Asymmetric substitution patterns in the two DNA strands of bacteria. *Mol Biol Evol* **13:** 660–665.

McInerney, J.O. (1998) Replicational and transcriptional selection on codon usage in *Borrelia burgdorferi*. *Proc Natl Acad Sci USA* **95:** 10698–10703.

Marians, K.J. (1992) Prokaryotic DNA replication. *Annu Rev Biochem* **61:** 673–719.

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

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

Salzberg, S.L., Salzberg, A.J., Kerlavage, A.R., and Tomb, J.-F. (1998) Skewed oligomers and origins of replication. *Gene* **217:** 57–67.

Smith, D.R., Doucette-Stamm, L.A., Deloughery, C., Lee, H., Dubois, J., *et al*. (1997) Complete genome sequence of *Methanobacterium thermoautotrophicum* ΔH: functional analysis and comparative genomics. *J Bacteriol* **179:** 7135–7155.

Stephens, R.S., Kalman, S., Lammel, C., Fan, J., Marathe, R., Aravind, L., *et al*. (1998) Genome sequence of an obligate intracellular pathogen of humans: *Chlamydia trachomatis*. *Science* **282:** 754–759.

Tomb, J.-F., White, O., Kerlavage, A.R., Clayton, R.A., Sutton, G.G., *et al*. (1997) The complete genome sequence of the gastric pathogen *Helicobacter pylori*. *Nature* **388:** 539–547.