G+C3 Structuring Along the Genome: A Common Feature in Prokaryotes

Vincent Daubin and Guy Perrière

Laboratoire de Biométrie et Biologie Évolutive, Université Claude Bernard — Lyon 1, Villeurbanne, France

The heterogeneity of gene nucleotide content in prokaryotic genomes is commonly interpreted as the result of three main phenomena: (1) genes undergo different selection pressures both during and after translation (affecting codon and amino acid choice); (2) genes undergo different mutational pressure whether they are on the leading or lagging strand; and (3) genes may have different phylogenetic origins as a result of lateral transfers. However, this view neglects the necessity of organizing genetic information on a chromosome that needs to be replicated and folded, which may add constraints to single gene evolution. As a consequence, genes are potentially subjected to different mutation and selection pressures, depending on their position in the genome. In this paper, we analyze the structuring of different codon usage measures along completely sequenced bacterial genomes. We show that most of them are highly structured, suggesting that genes have different base content, depending on their location on the chromosome. A peculiar pattern of genome structure, with a tendency toward an A+T-enrichment near the replication terminus, is found in most bacterial phyla and may reflect common chromosome constraints. Several species may have lost this pattern, probably because of genome rearrangements or integration of foreign DNA. We show that in several species, this enrichment is associated with an increase of evolutionary rate and we discuss the evolutionary implications of these results. We argue that structural constraints acting on the circular chromosome are not negligible and that this natural structuring of bacterial genomes may be a cause of overestimation in lateral gene transfer predictions using codon composition indices.

Introduction

Within a given bacterial species, genome fragments are relatively homogeneous for their G+C (Sueoka 1962), dinucleotides, trinucleotides, and longer oligonucleotides compositions (Karlin, Mràzek, and Campbell 1997; Karlin 1998; Wang 2001), this when compared with fragments of other species. It has been shown that codon usage is very biased in most species and often correlated with tRNA abundances, suggesting adaptation for a better translation efficiency (Ikemura 1981; Gouy and Gautier 1982; Bulmer 1987; Kanaya et al. 1999). On the other hand, codon usage is also influenced by mutation pressure. Hence, codon composition of a gene is the result of concurrent evolutionary forces that are differently balanced from one gene to another, depending on its function, expression rate, and position on the chromosome. This suggests that a combination of mutational and selection pressures typical to each genome tends to give it a unique vocabulary (Sueoka 1992; Karlin, Mràzek, and Campbell 1997). This finding led to the conclusion that searching for genes having atypical codon or base composition will allow identifying those that are adapted to another genomic context (i.e., "alien" genes). The underlying hypothesis here implies that a newly transferred gene will rapidly adapt its codon usage to its new genomic environment, and this approach is thus limited to recently acquired genes. Applied to several species, this methodology has been used to quantify the rate of gene transfer in bacteria. It gave, for some species, surprisingly large quantities of recently acquired genes (Médigue et al. 1991; Lawrence and Ochman 1998; Garcia-Vallvé, Romeu, and Pallau 2000; Ochman, Lawrence, and Groisman 2000), and the coherence of the genome concept has been consequently discussed (Bellgard et al. 1999; Doolittle 2000). However, evidence exists that other evolutionary pressures may have important effect on gene evolution. For example, it has been shown in enterobacteria that genes close to the replication origin are more conserved between species (Sharp et al. 1989). This view suggests that chromosomal location is a potential source of evolutionary and, consequently, gene content differences in prokaryotic genomes.

A first analysis performed on *Escherichia coli* genome already showed that the G+C composition at the third codon position (G+C3) varies along the genome in relation to the proximity of the replication terminus (Guindon and Perrière 2001). In the present work, we have completed and extended this analysis to 48 bacterial and 11 archaeal genomes. We show that the majority of bacterial and archaeal species display a significant structuring for some of the factors generally used to detect lateral transfers. Among bacterial structured genomes, two categories are identifiable: those concordant with the pattern observed in *E. coli*, which seem to have representatives in most bacterial phyla, and those showing mosaic structures that require other explanations involving lateral transfers and genome rearrangements. We also show that the evolutionary rates vary significantly along several bacterial genomes with a tendency for genes close to the replication terminus to evolve more rapidly. Taken together, these two observations suggest that a previously neglected evolutionary constraint may be responsible for A+T enrichment and increasing of evolutionary rates in a large chromosomal region around the replication terminus of bacteria.

Materials and Methods

G+C3, CAI and χ^2 Calculation

Complete genome sequences and their annotations were extracted from the EMGLib database (Perrière, Labedan, and Bessières 2000). After selection of coding sequences longer than 150 nucleotides, we computed the Codon Adaptation Index (CAI) (Sharp and Li 1987) and the frequency of G+C nucleotides at the third codon

Key words: complete genome, bacteria, mutation pressure, G+C content, lateral transfer, evolutionary rate.

E-mail: perriere@biomserv.univ-lyon1.fr.

Mol. Biol. Evol. 20(4):471–483. 2003
DOI: 10.1093/molbev/msg022

position (G+C3) for each gene. As the computing of CAI requires a reference table established using highly expressed genes, we calculated this table for each genome using the codon frequencies of all the genes coding for ribosomal proteins. Indeed, these genes are known to be highly expressed in unicellular organisms such as bacteria (Srivastava and Schlessinger 1990). We then calculated for each gene of length greater than 150 nucleotides the value of these parameters centered on the mean (denoted G+C3c and CAIc in this paper) and plotted the cumulative sum of these values along the genome (fig. 1). When the positions of the origin and terminus of replication were not available in the genome annotations, we determined them using the Oriloc program (Frank and Lobry 2000).

The amplitude of the cumulative sum graph directly depends on the structuring of the centered values. To test whether the amplitude observed was significantly different from a random distribution of genes, we resampled randomly 1,000 times the order of genes in the genome. We thus obtained a distribution of the amplitude of the cumulative sum graph under the hypothesis of a random repartition of genes in the genome. We considered that the structuring was significant ($P < 10^{-3}$) when the observed value was out of the range of this distribution.

Divergence Calculation

Divergence of gene sequences was estimated for several pairs of genomes (fig. 2). Homologous genes were identified using BLASTP2 (Altschul et al. 1997) searches between closely related genomes. Only proteins having E values less than 10^{-20} were considered as significant matches. Nucleotide sequences were then aligned with respect to the protein alignment. Ka and Ks values were calculated using JaDis (Gonçalves et al. 1999) and PAML (Yang 1997) for verification.

To be valid, the estimation of the effect of a gene position in the genome on its evolutionary rate must be investigated on genes having conserved their position relative to the replication origin and terminus in the two species compared. Pairs of closely related species for which complete genomes are available can be found in enterobacteria (*Salmonella* and *Escherichia*) and in the genera *Listeria*, *Neisseria*, *Rickettsia*, *Helicobacter*, *Chlamydia*, *Mycobacterium*, *Mycoplasma*, *Staphylococcus*, and *Streptococcus*. In archaea, three pairs of closely related species are available in the genera *Pyrococcus*, *Sulfolobus*, and *Thermoplasma*. However, in some genera, gene order is so poorly conserved that the results would not be interpretable. This is the case particularly for *Mycobacterium leprae* and *Mycobacterium tuberculosis*, *Streptococcus pneumoniae* and *Streptococcus pyogenes*, and *Sulfolobus solfataricus* and *Sulfolobus tokodaii*. Also, some pairs are so close that almost all their genes have identical nucleic sequences (e.g., the two strains of *S. pneumoniae* and the two strains of *E. coli* O157:H7). Thus, only eight pairs gave exploitable data: *Salmonella*/*Escherichia*, *E. coli* K12/*E. coli* O157:H7, *Listeria monocytogenes*/*Listeria innocua*, *Neisseria meningitidis* A/*Neisseria meningitidis* B, *Rickettsia prowazekii*/*Rickettsia conorii*, *Helicobacter pylori* J99/*Helicobacter pylori* 26695, *Chlamydia trachomatis*/*Chlamydia muridarum* and *Pyrococcus abyssi*/*Pyrococcus horikoshi*. We retained for statistical analysis only genes having conserved their position relative to the replication origin and terminus in both species. For this purpose, dot-plots of significant matches ($E < 10^{-20}$) between genomes were made as described in Eisen et al. (2000). Spearman's rho correlation nonparametric test was used for Ka and Ks values, and all statistics were computed using StatView™. For representation purpose (figs. 2 and 3), all the genomes have been arbitrary subdivided into seven groups of genes with respect to their distance to the replication origin.

Results

Unstructured Genomes

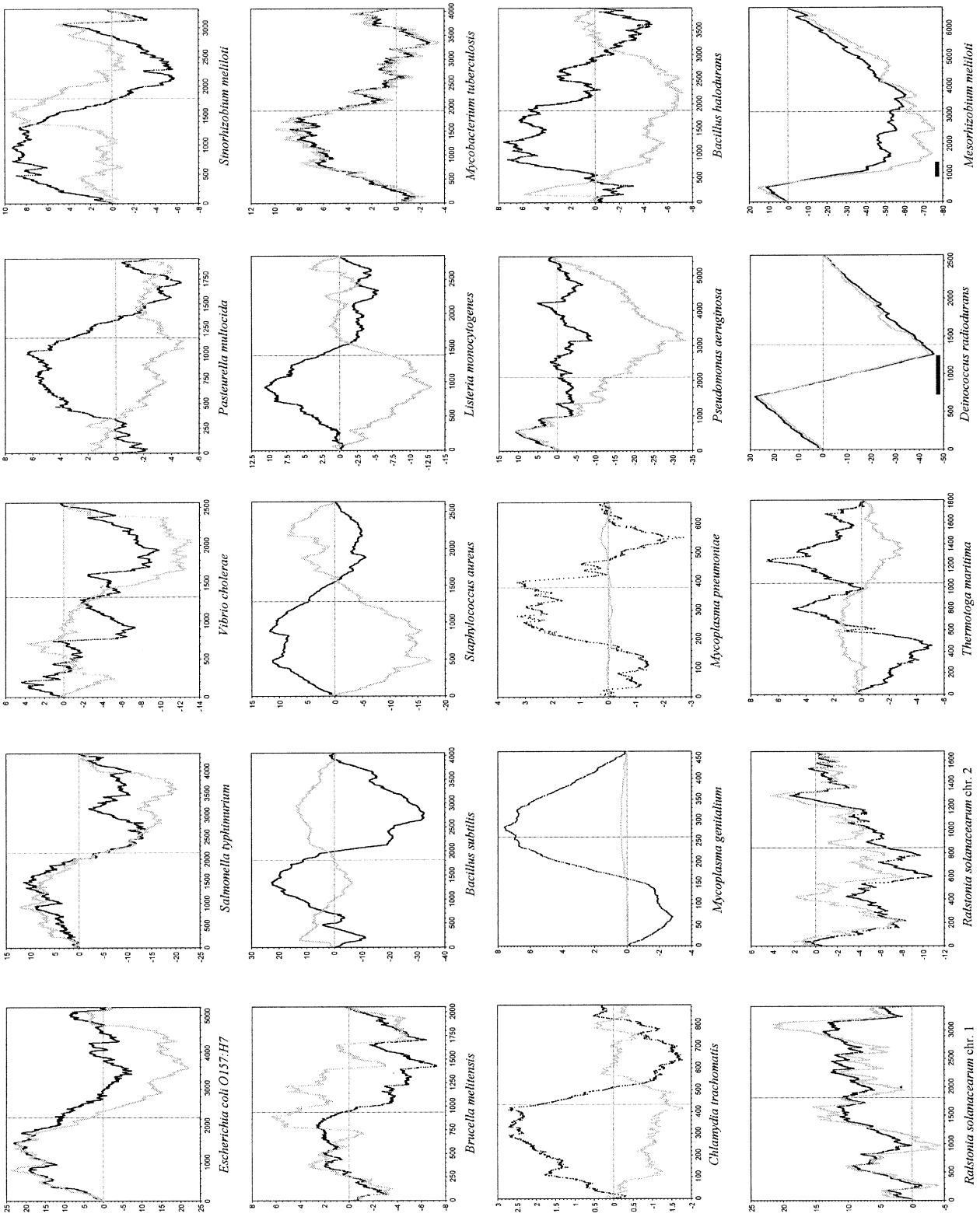
Among the 48 bacterial species tested for the structuring of the G+C3 along the chromosome, only six do not present a significant structure at $P < 10^{-3}$: *Nostoc* sp. PCC 7120, *Synechocystis* sp. PCC 6803, *Buchnera* sp. APS, *R. conorii*, *Borrelia burgdorferi*, and *Aquifex aeolicus* (table 1). These bacteria thus appear to be exceptions. Interestingly, three of them are known to be intracellular parasites and this lifestyle is known to have a strong impact on genomes due to relaxation of several adaptation pressures (Moran 1996). *R. conorii*, *Buchnera* sp., and *B. burgdorferi* have thus a very low G+C3 content and have undergone strong reduction of their genome size. However, the low number of genes and the A+T-richness of these genomes could not be the only explanation of their lack of structuring, because *Mycoplasma* species show very structured genomes. Since all of these species have become intracellular parasites independently, they may have retained different mechanisms to maintain the integrity of their genomes that may leave different marks.

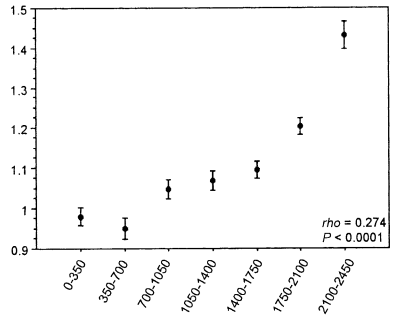
A special case is the genome of *B. burgdorferi*, which has very peculiar features since it is linear and has an extreme G+C skew between leading and lagging strands. This strand asymmetry is the major factor shaping its codon usage (McInerney 1998). Since the replication process necessarily undergoes very different constraints in this species, the difference of structuring compared with other bacteria is not surprising.

The two cyanobacteria (i.e., *Nostoc* sp. and *Synechocystis* sp.) only show a weak structuring ($P = 5.10^{-2}$), while the hyperthermophilic bacteria *A. aeolicus* does not show a significant structuring of its genome. Their case, and especially the one of *A. aeolicus*, is very puzzling, and we lack information on the replication process in these

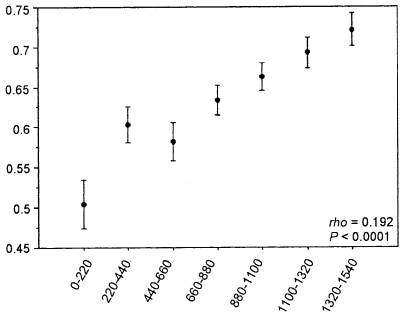
→

FIG. 1.—Cumulative sums for G+C3c (black curve) and CAIc (gray curve). Horizontal axes: cumulative sum value; vertical axes: position on the chromosome. Vertical dotted bars show the position of the replication terminus (see text). Graphs for χ^2c are not shown for clarity but globally follow the CAIc curves. Only 20 representative species are shown. Black boxes show the regions of homology with plasmids in the genomes of *D. radiodurans* and *M. meliloti*.

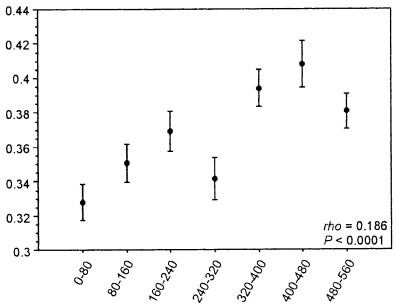

Escherichia coli O157:H7
Salmonella typhimurium
Vibrio cholerae
Pasteurella multocida
Sinorhizobium meliloti
Brucella melitensis
Bacillus subtilis
Staphylococcus aureus
Listeria monocytogenes
Mycobacterium tuberculosis
Chlamydia trachomatis
Mycoplasma genitalium
Mycoplasma pneumoniae
Pseudomonas aeruginosa
Bacillus halodurans
Ralstonia solanacearum chr. 1
Ralstonia solanacearum chr. 2
Thermotoga maritima
Deinococcus radiodurans
Mesorhizobium meliloti



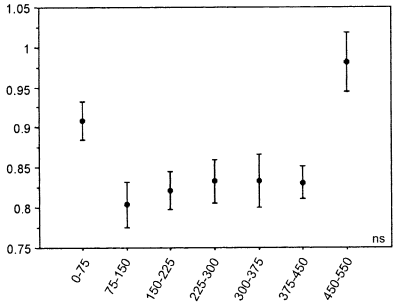
Ks *Escherichia coli* K12 - *Salmonella typhimurium* LT2



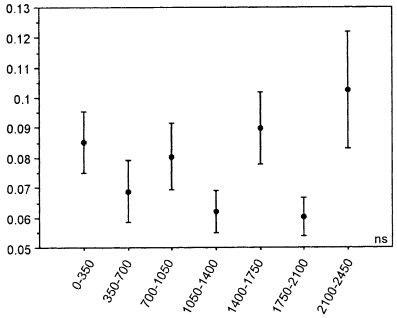
Ks *Listeria monocytogenes* - *Listeria innocua*



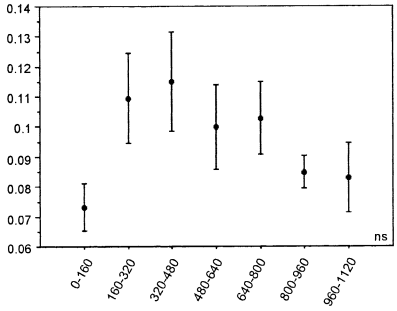
Ks *Rickettsia prowazekii* - *Rickettsia conorii*



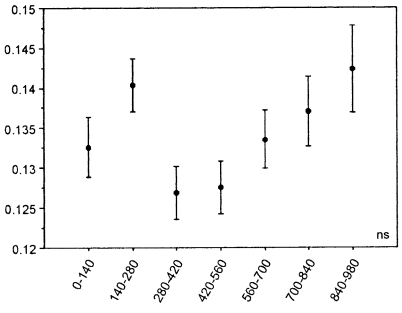
Ks *Chlamydia trachomatis* - *Chlamydia muridarum*



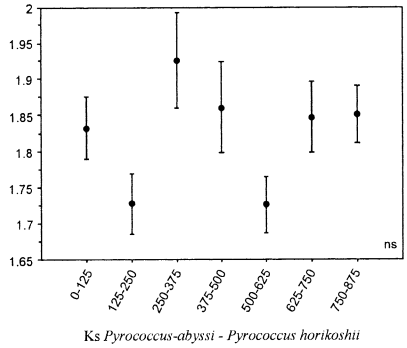
Ks *Escherichia coli* K12 - *E. coli 0157:H7*



Ks *Neisseria meningitidis* A - *N. meningitidis* B



Ks *Helicobacter pylori* J99 - *H. pylori* 26695



Ks *Pyrococcus-abyssi* - *Pyrococcus horikoshii*

FIG. 2.—Relation between distance to the replication origin and synonymous evolutionary rate. Horizontal axes: distance intervals from the origin in kb; vertical axes: number of synonymous substitution per site. Each genome has been arbitrary divided into seven parts with respect to the distance to the origin. Only genes having conserved their position in both organisms have been considered. The mean values and the standard error is shown for each part. Significance and rho values for the Spearman's test are shown.

bacteria to understand this peculiar feature. Interestingly, contrarily to most bacterial species, *Synechocystis*, *Nostoc*, and *Aquifex* do not present the very common G+C skew between strands allowing prediction of the replication origin (Lobry 1996; Karlin, Campbell, and Mràzek 1998; Lopez et al. 1999). This suggests that their replication mechanisms may present some important differences from other bacteria. Since the position of the replication terminus is still unknown in cyanobacteria, it was impossible to link the weak structuring with replication landmarks.

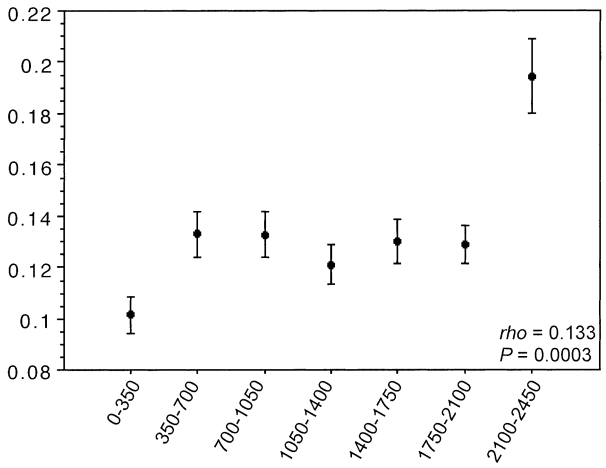
The Structuring of G+C3: Variations on a Common Theme

All other bacteria or archaea tested showed a strong structuring for their G+C3 at the scale of their complete genome. Among them, we find species from most large taxonomic groups of bacteria: low G+C gram positives, high G+C gram positives, Proteobacteria, Chlamydiales, Thermotogales, Spirochaetes, and Deinococcales (table 1). Only Aquificales and Cyanobacteria do not present evidence of such a structuring, possibly because of a sampling effect, since only a few bacteria of these phyla have been presently sequenced. Several species showing a structuring are endoparasites (like Chlamydiales), suggesting that the parasitic way of life is not a sufficient reason for the lack of structuring observed in other species. The structuring of the G+C3 is often accompanied by a structuring of the CAI. This index can be either positively or negatively correlated to G+C3, depending on the optimal codons (G+C-rich or A+T-rich) used in the species (fig. 1). The use of other codon usage indices such as χ^2-weighted by length (Shields and Sharp 1987) yields results very similar to CAI structuring (results not shown).

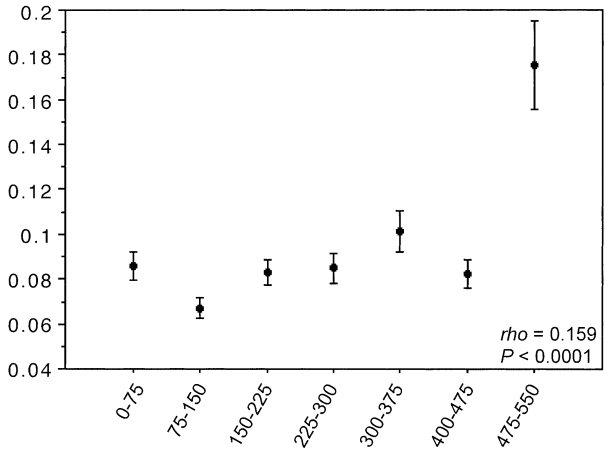
Two types of genome structuring were identified in our analysis. The first type corresponds to the one described by Guindon and Perrière (2001) in *E. coli* K12. It is characterized by enrichment in A+T near the replication terminus of the chromosome. Bacteria phylogenetically close to *E. coli* K12, such as the pathogenic strain O157:H7 and the two *Salmonella* species, show very similar patterns. However, bacteria as diverse as *Pasteurella multocida*, *Sinorhizobium meliloti*, *Brucella melitensis*, *R. prowazeki*, *C. trachomatis*, *Bacillus subtilis*, *Staphylococcus aureus*, *L. monocytogenes*, and *M. tuberculosis* have the same typical A+T enrichment near the terminus region. The G+C content of intergenic regions follows exactly the same pattern (results not shown). As well, when considering only genes having conserved their position relative to the origin and terminus between *E. coli* and *S. typhimurium* (i.e., genes already present in the common ancestor), the same pattern is found (fig. 4).

The second type of genome structure is more complex and suggests a mosaic organization of the chromosome. This is the case for several species, such as *Pseudomonas aeruginosa*, *Ralstonia solanacearum*, *Bacillus halodurans*, *Thermotoga maritima*, and *Mycoplasma pneumoniae*.

Since the structuring is very common in bacteria, we also investigated the case of archaeal species. All of the 11 species tested present several regions of contrasting G+C content. However, although evidence suggests that archaea, like bacteria, possess a unique origin from which replication initiates bidirectionally (Lopez et al. 1999; Myllykallio et al. 2000; McNeill 2001), the majority of archaeal species still have no confirmed position for their replication origin and terminus. Recently, the position of the origin of *P. horikoshii*, *P. abyssi*, *Methanobacterium thermoautotrophicum*, and *S. solfataricus* have been predicted using cumulative skew (Lopez et al. 1999; She et al. 2001) and experimentally confirmed for one of these species (*P. abyssi*) (Myllykallio et al. 2000). The identified origins always coincide with a *cdc6*/*orc1* locus, which is known to play a role in eukaryotic replication. In *Pyrococcus* and *Sulfolobus*, if we suppose that the terminus is located oppositely from the origin on the circular chromosome, the corresponding region is an A+T-rich one, although relatively short compared with the analogous regions observed in bacteria (fig. 1). *Methanobacterium*, on the other hand, do not show such



Ka *Escherichia coli* K12 - *Salmonella typhimurium* LT2



Ka *Chlamydia trachomatis* - *Chlamydia muridarum*

FIG. 3.—Relation between distance to the replication origin and nonsynonymous evolutionary rate. The data have been treated as described in figure 2.

Table 1
List of the 59 Species Tested

Phylum	Species Name	Gene#	$Mean_{G+C3}$	Str_{G+C3}	Str_{CAI}	Ter
Bacteria						
Gram positive						
	Bacillus halodurans[a]	3950	42,1	+++	+++	n
	Bacillus subtilis[a]	4052	43,6	+++	+++	y
	Staphylococcus aureus[a]	2638	23,0	+++	+++	y
	Streptococcus pneumoniae[a]	2015	35,6	+++	+++	y
	Streptococcus pyogenes[a]	1682	31,7	++	+++	n
	Clostridium acetobutylicum[a]	3651	21,3	+++	+++	y
	Lactococcus lactis subsp. *lactis*[a]	2257	25,6	+++	+++	n
	Listeria innocua[a]	2969	28,9	+++	+++	y
	Listeria monocytogenes[a]	2849	30,0	+++	+++	y
	Mycoplasma genitalium[a]	466	23,3	+++	+++	n
	Mycoplasma pneumoniae[a]	674	41,1	+++	+	n
	Mycoplasma pulmonis[a]	774	15,3	+++	++	n
	Ureaplasma urealyticum[a]	607	12,9	+++	++	n
	Mycobacterium leprae[b]	2691	49,6	++	+++	n
	Mycobacterium tuberculosis[b]	4062	78,1	+++	+++	y
Cyanobacteria						
	Nostoc sp. PCC 7120	5329	35,2	+	−	?
	Synechocystis sp. PCC 6803	3103	49,6	+	+	?
Proteobacteria						
	Escherichia coli O157:H7[c]	5208	53,6	+++	+++	y
	Escherichia coli K12[c]	4254	54,5	+++	+++	y
	Salmonella enterica[c]	4519	56,2	+++	+++	y
	Salmonella typhimurium[c]	4401	57,9	+++	++	y
	Buchnera sp. APS[c]	562	14,4	−	−	−
	Vibrio cholerae[c]	2562	48,6	+++	++	y
	Haemophilus influenzae[c]	1647	29,0	+++	++	n
	Pseudomonas aeruginosa[c]	5551	86,7	+++	++	n
	Xylella fastidiosa[c]	2645	55,3	+++	++	y
	Yersinia pestis[c]	3976	47,9	+++	++	y
	Pasteurella multocida[c]	2011	34,4	+++	++	y
	Neisseria meningitidis[d]	2065	58,7	+++	+	n
	Ralstonia solanacearum[d]	3417	86,3	+++	+++	n
	Sinorhizobium meliloti[e]	3326	78,8	+++	+++	y
	Mesorhizobium loti[e]	6705	78,7	+++	+++	n
	Brucella melitensis[e]	2055	65,9	+++	+++	y
	Agrobacterium tumefaciens[e]	2679	71,6	+++	−	n
	Rickettsia conorii[e]	1372	23,5	−	−	−
	Rickettsia prowazekii[e]	830	18,4	++	+	y
	Caulobacter crescentus[e]	3684	85,5	+++	+++	y
	Campylobacter jejuni[f]	1620	19,5	+++	+++	n
	Helicobacter pylori J99[f]	1477	42,2	+++	+	n
	Helicobacter pylori[f]	1513	41,5	+++	−	n
Chlamydiales						
	Chlamydia muridarum	797	33,5	+++	+	y
	Chlamydophila pneumoniae AR39	941	34,5	+++	+++	y
	Chlamydia trachomatis	891	34,6	+++	−	y
Spirochaetes .						
	Borrelia burgdorferi	821	20,9	−	−	−
	Treponema pallidum	1000	54,8	++	+++	y
Thermotogales						
	Thermotoga maritima	1810	52,3	+++	+++	n
Aquificales . .						
	Aquifex aeolicus	1522	47,9	−	++	−
Deinococcales						
	Deinococcus radiodurans	2577	79,9	+++	++	n
Archaea						
Crenarchaeota						
	Aeropyrum pernix	2694	65,3	+++	+++	?
	Sulfolobus solfataricus	2971	33,3	+++	+++	y (?)
	Sulfolobus tokodaii	2826	25,6	+++	+++	?
Euryarchaeota						
	Halobacterium sp. NRC-1	2017	87,3	+++	+++	?
	Methanococcus jannaschii	1674	27,7	++	+++	?
	Methanobacterium thermoautotrophicum	1859	55,9	+++	+++	n
	Pyrococcus abyssi	1764	50,2	+++	+++	y (?)
	Pyrococcus horikoshi	1979	42,9	+++	+++	y (?)

Table 1
Continued

Phylum	Species Name	Gene#	$Mean_{G+C3}$	Str_{G+C3}	Str_{CAI}	Ter
	Thermoplasma acidophilum	1477	54,1	+++	+++	?
	Thermoplasma volcanium	1495	41,0	+++	+++	?
	Archaeoglobus fulgidus	2374	57,8	+++	+++	?

NOTE.—The results for the χ^2c structuring are not shown explicitly but are globally the same as for CAIc. +++ indicates test significant at $P < 10^{-3}$; ++ indicates test significant at $P < 10^{-2}$; + indicates test significant at $P < 5.10^{-2}$; − indicates test not significant. In the case of bacteria containing more than one chromosome (*A. tumefasciens*, *B. melitensis*, *D. radiodurans*, *Nostoc* sp. PCC 7120, *R. solanacearum*, and *V. cholerae*), we only indicate values for the largest one if the results are not different. Ter_{A+T} indicates whether the structuring involves an A+T enrichment near the terminus (y) or not (n). ? indicates that the position of the replication terminus is uncertain. For *S. pneumoniae*, *N. meningitidis*, and *C. pneumoniae*, only the results for one strain are given since they are very similar. Gene# is the number of genes longer than 150 nucleotides. $Mean_{G+C3}$ is the G+C3 mean. Str_{G+C3} is the result of the test for the G+C3 structuring. Str_{CAI} is the result of the test for the CAI structuring.

[a] Low G+C.
[b] High G+C.
[c] γ-Proteobacteria.
[d] β-Proteobacteria.
[e] α-Proteobacteria.
[f] ε-Proteobacteria.

patterns. However, this result must be confirmed since no clear archaeal terminus has yet been identified. In other archaeal species, placing the replication terminus seems even more risky, since it has been proposed for example that *Halobacterium* species could have several replication origins (Ng et al. 2000).

Evolutionary Rate Variation Along the Genome

Sharp et al. (1989) have observed that evolutionary rates tend to increase with the distance to the replication origin in enterobacteria. Since their work was based on a few genes, we have repeated the experiment using genes having conserved their distance to the replication origin between *E. coli* K12 and *Salmonella typhymurium* and the other pairs listed in the *Material and Methods* section.

The results are shown in figure 2. In three of the seven pairs tested, we found a significant increase of the synonymous evolutionary rate (Ks) with the distance to the replication origin. The same is observed for Ka (fig. 3) in *Chlamydia* and enterobacteria. *Chlamydia*, *Neisseria*, *Helicobacter*, and *Pyrococcus* pairs do not show a significant tendency for Ks increase with the distance to the replication origin. It is interesting to note that genomes showing a clear A+T enrichment near the terminus region (i.e., enterobacteria, *Rickettsia*, *Listeria*, and *Chlamydia*) also display structuring of at least one of the two evolutionary rates tested. However, *Neisseria* and *Helicobacter* strains have diverged recently as witnessed by their low Ks values. The comparison of *E. coli* K12 with both *E. coli* O157:H7 and *S. typhymurium* shows that the effect of the distance to the origin needs a divergence time large enough to be observed. Thus, one cannot exclude that the low increase of evolutionary rate with distance from the origin in *Helicobacter* could become significant with time. *Neisseria* species do not show such tendency. Remarkably, the most divergent pairs (i.e., *Chlamydia* and enterobacteria) also present an increase of Ka with the distance to the origin (fig. 3). This is particularly surprising for the *Chlamydia* pair since it shows no significant increase of the Ks. However, the shape of the graph suggests that it may be due to a peculiar set of genes having high synonymous evolutionary rate in the region of the origin in these species.

These results confirm those obtained by Sharp et al. (1989) in *E. coli* and extend their observation to some other bacterial species. Also, it shows different patterns of evolutionary rate increases along the chromosome (i.e., rather linear for *Rickettsia* and *Listeria* or exponential for enterobacteria and *Chlamydia*) from one bacterium to another. This suggests possible specific effects of replication origin and terminus proximity on evolutionary rate, depending on the organisms considered. However, it remains that, although the increase with the distance to the origin is not always significant, five of the six genera tested show particularly high mean evolutionary rate for genes close to the replication terminus compared with other regions of the genome.

Discussion

Similarities and Differences in G+C Structuring

Given the widespread of the G+C3 pattern first observed in *E. coli* into bacterial phyla, we propose that

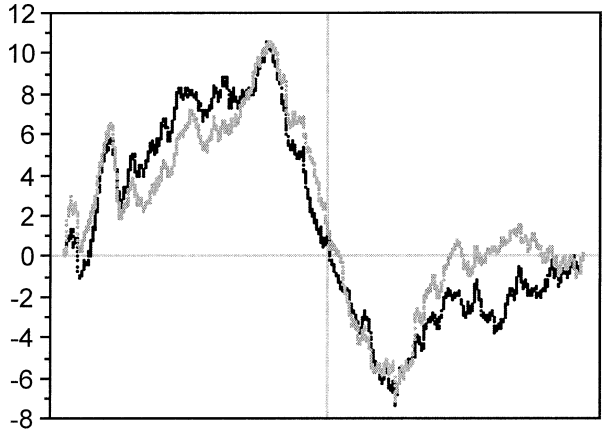


FIG. 4.—Cumulative sums for G+C3c for *E. coli* (black curve) and *S. typhimurium* (gray curve) when considering only genes conserved from their common ancestor.

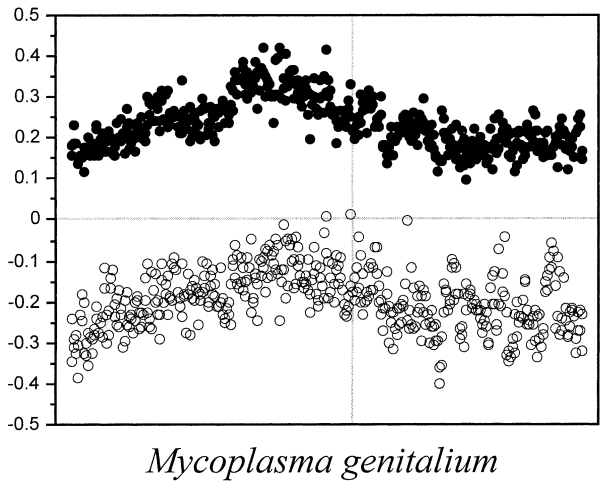


FIG. 5.—G+C3 (black dots) of genes along the genomes of *M. genitalium* and difference between G+C3 in *M. genitalium* and *M. pneumoniae* ($G+C3_{Mg} - G+C3_{Mp}$) (circles). The difference is always negative, which is consistent with the A+T enrichment of the *M. genitalium* genome, but this difference is much stronger in the more A+T-rich region, suggesting a recent acquisition of this pattern.

this genome structuring rests on a mechanism that is common to every bacteria, and that the second type of structures is derived from the first either by genome reorganization or import of foreign DNA. The mosaic structure sometimes corresponds clearly to a single event of lateral transfer or chromosome fusion. For example, the first chromosome of the radio-resistant bacteria *Deinococcus radiodurans* (fig. 1) displays a striking highly structured pattern, which is due to the probable fusion of this chromosome with a copy of the megaplasmid also present in the bacteria. This results in two regions of the chromosome having homogeneous G+C3 content. The first one corresponds to the probable original single chromosome, as it carries most orthologous genes present in the closely related bacteria *Thermus*. The second one, possessing several homologues on the megaplasmid (White et al. 1999), has a much lower G+C3 content. The high difference of G+C3 content between these two segments could be explained either by the recent fusion or by concerted evolution due to frequent recombination events between the megaplasmid and its homologous segment on the chromosome.

The pattern of the first chromosome of *Mesorhizobium loti* (fig. 1) seems to be due to the same kind of event since the region of low G+C3 content contains homologs of a plasmid present in closely related α-Proteobacteria (Kaneko et al. 2001). However, for the other bacteria, the reason for the nonrandom repartition of genes with respect to their G+C3 remains under question.

The case of the two close species *Mycoplasma genitalium* and *M. pneumoniae* is particularly interesting. The first one shows a very intriguing profile already noted by Kerr, Peden, and Sharp (1997). There is indeed a peculiar variation of the G+C content along the chromosome, but contrarily to most bacteria showing a similar pattern, it is clearly not related to the position of the replication origin and terminus (Kerr, Peden, and Sharp 1997). Surprisingly, the origin takes place in a rather A+T-rich region (18.4% G+C3), but the terminus is in a region where the G+C content is close to the mean of the genome (23.3% G+C3). The second has a very perturbed pattern. Though they both are obligate parasites, it seems that the dependence of *M. genitalium* on its host is stronger, since it is much more difficult to grow in artificial media (Jensen, Hansen, and Lind 1996). Interestingly, *M. genitalium* has a more reduced genome and a lower G+C3 content than *M. pneumoniae*. However, Himmelreich et al. (1997) have compared the two genomes and showed that the gene order was highly conserved within six fragments, concluding that these rearrangements occurred in the lineage of *M. genitalium*. The rearrangements do not correspond, contrarily to what is observed in most bacterial genomes (Eisen et al. 2000), to inversions around the origin and terminus. Himmelreich et al. (1997) have shown that these rearrangement events are due to the presence of repeats that are completely conserved only in *M. pneumoniae*. It has been proposed that rearrangements play a significant role in these bacteria for the mechanism of virulence (Rocha and Blanchard 2002). Himmelreich et al. (1997) have proposed that the rearrangements events observed here took place in *M. genitalium*. This suggests that the regular pattern found in this species has shaped rapidly after the speciation of the two *Mycoplasma* organisms, simultaneously with the reduction in G+C content. The analysis of the differences in G+C3 between the two species indeed suggests clearly that the reduction in G+C content has been much stronger in the A+T-rich region of the *M. genitalium* genome (fig. 5). The mechanism explaining this feature remains mysterious. However, the peculiar structuring as well as the rearrangement pattern suggests that the replication mechanism in *M. genitalium* possesses atypical properties in bacteria.

The two chromosomes of the plant pathogenic bacteria *R. solanacearum* are both strongly structured but also with a pattern suggesting a mosaic. This bacterium is known to be able to acquire foreign DNA by natural transformation (Bertolla et al. 1997), and genomic rearrangements have been reported to occur naturally in its genome (Brumbley, Carney, and Denny 1993). Salanoubat et al. (2001) already described this heterogeneity, with large regions of different G+C content and codon usage in the complete genome, and attributed this mosaïc structure to recent gene transfers.

Evolutionary Rate Heterogeneity: Differential Selection or Mutation?

The observed increasing evolutionary rate and A+T-richness in the region of the terminus can be explained by weaker selection efficiency in this region or a higher mutational pressure. Since bacterial codon usage is more or less biased, the Ks heterogeneity could be due to differences in codon usage bias among genes. Thus, a biased location of highly expressed genes would explain our results. However, Sharp et al. (1989, p. 809) noted that there was "no significant linear relation between codon bias and distance from the origin, reflecting the fact that highly expressed genes are not predominantly clustered

near OriC." They concluded that map position has no effect on selective constraints in enterobacteria. Although the structuring of CAI suggests an opposite explanation, the more reproducible shape of G+C3c cumulative curves shows that constraints on G+C content are more likely responsible for this. In several species, including enterobacteria, *B. melitensis*, *Vibrio cholerae*, *S. meliloti*, and *M. tuberculosis*, in which the CAI is positively correlated to the G+C3, CAI values for the genes near the terminus are lower than in the remainder of the chromosome. In contrast, this trend is inverted for species in which optimal codons are A+T-rich (e.g., *B. subtilis*, *L. monocytogenes*, and *S. aureus*). Thus, the main explanation for CAI structuring is clearly the G+C3 content. Indeed, favoring a structuring of highly expressed genes would require hypothesizing an opposite strategy of gene organization in A+T-rich and G+C-rich genomes. Indeed, in a genome, genes may display high CAI values because they undergo high selection on their codon usage or because the mutational bias enriches them in preferred codons. It seems probable that in A+T-rich genomes, the high values of CAI in the terminus region are due to the second mechanism. Conversely, it is interesting to note that some bacteria (i.e., *P. aeruginosa* or *T. maritima*), which do not present a structuring of G+C3 related to the position of the terminus, display a strong decrease of CAI values in this region.

To test whether the observed increasing of Ks values can be attributed to a clustering of highly expressed genes in the origin region, we have divided genes having low and high CAI values and found that the increasing of Ks was significant ($P < 10^{-4}$) within each class for enterobacteria, *Listeria*, and *Rickettsia*. This excludes the hypothesis of a pattern due to higher codon bias in the origin region and confirms the existence of differential mutation rates along genomes of several bacterial species. The particularly high values of Ks in the terminus region of the four pairs showing an A+T-enrichment of this region (i.e., enterobacteria, *Listeria*, *Rickettsia*, and *Chlamydia*) suggests that this region is subject to particularly high mutation rates. This is comforted by the Ka variations for enterobacteria and *Chlamydia* (fig. 3), since every region of the chromosome display similar Ka values except this region. In *Chlamydia*, the increase of Ks with distance to the origin is not significant, probably because of the high values found near the origin.

Why Peculiar Patterns for Genes Near the Terminus?

Several hypotheses may explain the pattern of decreasing G+C3 content observed near the replication terminus. One may consider, for instance, that the low G+C3 content of this region is due to a higher frequency of insertions of laterally transferred genes in this region. Indeed, Lawrence and Ochman (1998) have found that the genes they predicted as laterally transferred in *E. coli* were preferentially located in this region and had a low G+C content. Moreover, in the case of *B. subtilis*, several prophages, including a large one (SPβ) are known to be inserted in the terminus region (Kunst et al. 1997) and confers to the G+C3c cumulative curve its particularly high slope. However, even when removing the sequences from these prophages, the decrease in G+C3 content is still visible. Furthermore, when considering only genes conserved between *E. coli* and *S. typhimurium*, the same pattern is observed in both species (fig. 4). It seems surprising that genes transferred to the common ancestor of these enterobacteria have retained such a strong bias of nucleotide composition, despite the particularly high Ks values observed in this region. This rather suggests that an intrinsic enrichment of this region could be responsible for a misleading of methods based on nucleotide content to predict recently acquired genes. Moreover, the case of *C. trachomatis* hardly fit with the hypothesis of alien genes since few or no candidates for recently acquired genes have been found in this species (Garcia-Vallvé, Romeu, and Pallau 2000; Ochman, Lawrence, and Groisman 2000).

Another hypothesis is that structural constraints due either to peculiar supercoiling of the terminus region or to the resolution of chromosome dimers at the end of the replication gives this region a peculiar composition. First, Capiaux et al. (2001) have shown that some oligomers tend to increase in frequency near the terminus in *E. coli*, suggesting peculiar constraints acting in this region. Also, Ussery et al. (2001) have found that Fis, a very abundant and pleiotropic architectural protein involved in the structure of bacterial chromatin and the regulation of several genes (Finkel and Johnson 1992; Schneider et al. 2001; Travers, Schneider, and Muskhelishvili 2001), possesses a high density of binding sites in a large region of about 1 Mb around the *E. coli* terminus corresponding to the A+T-rich part of the chromosome. This region has also been found to be highly enriched in sequences favoring DNA curvature in both *E. coli* and *B. subtilis* (Pedersen et al. 2000), which may favor the fixation of other proteins involved in chromosome condensation, such as H-NS in *E. coli* (Ussery et al. 2001). The resulting structure of this large region encompassing terminus sites may play a role in the segregation of neosynthesized chromosomes in *E. coli* (Tsai and Sun 2001) and/or the resolution of chromosome dimers at the *dif* site. These processes have been shown to be highly dependent on flanking sequences (Perals et al. 2000). The mechanisms of chromosome replication termination in *E. coli* and *B. subtilis* are thought to have evolved independently (Hill 1992; Wake 1997) but occur via extremely similar mechanisms (Wake 1997; Bussière and Bastia 1999). Although no homologue of *fis* has been found in *B. subtilis*, a protein, AbrB, seems to share very similar characteristics in terms of size, DNA binding, expression pattern, and control on gene expression, which suggests that it could play an equivalent role (O'Reilly and Devine 1997). Thus, the A+T richness of the terminus region could have a functional interest for the replication process by facilitating protein binding and loop formation at least in *E. coli* and *B. subtilis*. Hence, the peculiar characteristics of this region would reveal a conflict between different levels of selection (i.e., at the gene level and at the chromosome level). Relatively little is known about the termination mechanism in other bacterial and archaeal species. Though the *ter* sites—which inhibit the action of

helicases ahead of the fork in a polar manner, precluding the replication forks to exit the terminus region—are believed to play a significant evolutionary role, it is worth noting that their deletion in *E. coli* and *B. subtilis* have no detectable effect on fitness in laboratory conditions (Bierne and Michel 1994). This suggests that the mechanism is dispensable and that some species might not possess such systems.

Several constraints apply at the replication terminus: (1) the two forks have to meet simultaneously at the *dif* site, sometimes necessitating the arrest of one of the forks in the terminus region by *ter* sites (Wake 1997; Bussière and Bastia 1999); (2) structural constraints may arise from the meeting of these forks (Lewis 2001); (3) chromosomes catenates and dimers have to be resolved (Lemon, Kuster, and Grossman 2001; Lewis 2001; Perals et al. 2001); and (4) the terminus region may also play a role in the segregation of chromosomes notably by interacting with XerCD and FtsK (Perals et al. 2001). These different constraints could be responsible both for an A+T-enrichment and for an enhancement of genes mutation rates. For example, the presence of DNA ends and persistent single-stranded DNA regions characterize an arrested replication fork. Therefore, it may be more sensitive to mutation or recombination processes (Bierne, Ehrlich, and Michel 1997).

On the other hand, the terminus region may also differ in its base composition due to the preferential use of specific repair systems. Sharp et al. (1989) have proposed such a hypothesis to explain the correlation between the distance between the replication terminus and the evolutionary rate in enterobacteria. They hypothesized that the presence of multiple forks during chromosome replication allows more frequent recombination events in the region of the replication origin and consequently provides a more efficient mutation repair. Under this model, sequences close to the replication terminus are in single copy for a longer part of the cell cycle than are the origin-linked genes, so they have fewer opportunities to engage in repair via homologous recombination. However, this hypothesis hardly fit with recent observations showing that the origins are segregated at opposites poles of the cell during replication, precluding contacts between these sequences (Sawitzke and Austin 2001). We rather propose another possible difference between the origin and terminus in replication-correlated repair mechanisms. The region of the replication terminus in *E. coli* contains at least 10 *ter* sites that, by combining to Tus proteins, inhibit the action of helicases ahead of the replication complex in a polar manner (Bussière and Bastia 1999). This allows the forks to meet close to the *dif* site, where chromosome dimers and catenates are resolved. When the replication complex meets a DNA lesion, the fork is stalled. This lesion must be repaired or bypassed by the replication machinery (Cox et al. 2000). This requires a regression of the fork, that is, the melting of the neosynthesized strands from their matrices and the formation of a Holliday junction through the action of the specific helicases RecG and PriA in a direction opposite to replication (McGlynn and Lloyd 2001; Gregg et al. 2002). However, it is possible that after the fork has entered the terminus region, the presence of *ter*/Tus complexes precludes this fork regression by inhibiting the PriA and/or RecG helicases. These complexes have indeed been shown to inhibit the action of PriA in vitro, although their action on RegG is still unknown (Hiasa and Mirans 1992). The DNA synthesis has then to be completed by polymerases from the translesion pathway. This mechanism is both error-prone and biased toward A+T nucleotides. The polymerases involved in the SOS system (notably polII and polV), which undertake "translesion" synthesis, are known to follow the "A-rule" (Strauss 1991; Ide et al. 1995), that is, the preferential incorporation of a dAMP at abasic sites. Since these lesions can arise from several spontaneous and induced mechanisms and are therefore very frequent (Ide et al. 1995), the A-rule could be responsible for the A+T enrichment of a region lacking recombination-dependent repair.

The mechanism proposed above might seem to be in contradiction with recent work by Hudson et al. (2002). These authors have inserted a nonfunctional *lacZ* gene in different locations of the *Salmonella enterica* chromosome and measured the frequency of reversion. They found no significant difference in reversion rates between genes inserted near the replication origin and genes inserted near the terminus. They examined several types of substitutions, including transition and transversion. However, it is interesting to note that all the possible reversions were mutation from A or T to C or G nucleotides. Thus, if the rate of mutations toward A+T increases along the chromosome, this study would not have detected it.

Implications for Lateral Transfer Detection

As shown by Ragan (2001), different intrinsic methods of lateral transfer detection fail to give consistent results in *E. coli*. We observe a structuring for CAI and χ^2 values along bacterial chromosomes, which shows that different parts of the genome may have different codon usage. The pattern observed in *E. coli*, which might be the result of a replication correlated mechanism, seems to apply to virtually every bacterial phylum. Thus, the underlying hypothesis of laterally transferred genes detection methods (i.e., the weak intrinsic heterogeneity of base content among genes of a genome) might not be applicable for most species. Although the lateral transfer of large DNA fragments remains a valuable hypothesis for explaining the structuring observed in some genomes, the natural tendency of genomes to be structured is a potential source of overestimation of alien genes. This implies potential bias on codon composition approaches, which are often used to detect alien genes (Karlin, Campbell, and Mràzek 1998; Lawrence and Ochman 1998; Garcia-Vallvé, Romeu, and Pallau 2000; Ochman, Lawrence, and Groisman 2000). Especially, it seems unjustified to suppose *a priori* that the G+C3 of genes follows a normal distribution in a genome (Lawrence and Ochman 1998). As we already mentioned, Lawrence and Ochman (1997, 1998) noted in their work that the hypothetically transferred genes they detected were significantly more represented in the terminus region. However, this possible overestimation does not only concern species in which the

genome presents a structuring resembling the *E. coli* genome. Indeed, rearrangement of a structured genome will still produce a structured genome since it will take a lot of events to randomize the distribution of genes.

Conclusion

It is now well known that both G+C content and evolutionary rates of genes are under the control of several factors, including gene expression level, codon bias, protein hydrophobicity, and strand location. We have presented here evidence that another factor may contribute to gene heterogeneity in several bacterial genomes: the proximity to the replication terminus. This may correspond to constraints at the chromosome level. We retained two hypotheses that may explain the peculiar features of the region of the replication terminus: (1) processes that may lead to a conflict between different levels of selection (i.e., the gene level and the chromosome level) in this region and (2) constraints due to a difference in mutation frequency and/or repair mechanism. These two hypotheses are not mutually exclusive and may both play a role. Interestingly, a few species do not present any structuring of their G+C content. A better understanding of the replication process in these species may highlight their peculiarities and allow a choice between these hypotheses.

The tendency for the terminus region to undergo mutational pressure biased toward A+T nucleotides could be ultimately demonstrated by a parsimony analysis of sets of three completely sequences genomes. This would allow orientating mutations with respect to an outgroup and testing whether mutation toward A+T nucleotides are more frequent in the terminus region. However, such an analysis is not presently possible with available data, because, for instance, *E. coli* 0157:H7 strains are too closely related, and Ks values currently exceed 1 between *E. coli* and *Salmonella*, precluding a parsimony hypothesis. New completely sequenced genomes should allow testing of our model of genome evolution.

Acknowledgments

We thank E. Lerat for daily discussions and L. Duret for critical reading of the paper. This work has been supported by CNRS. V.D. is recipient of a fellowship from the Ministère de l'Education Nationale de la Recherche et de la Technologie.

Literature Cited

Altschul, S. F., T. L. Madden, A. A. Schäffer, J. Zhang, Z. Zhang, W. Miller, and D. J. Lipman. 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. **25**:3389–3402.

Bellgard, M. I., T. Itoh, H. Watanabe, T. Imanishi, and T. Gojobori. 1999. Dynamic evolution of genomes and the concept of genome space. Ann. NY Acad. Sci. **870**:293–300.

Bertolla, F., F. Van Gijsegem, X. Nesme, and P. Simonet. 1997. Conditions for natural transformation of *Ralstonia solanacearum*. Appl. Env. Microbiol. **63**:4965–4968.

Bierne, H., S. D. Ehrlich, and B. Michel. 1997. Deletions at stalled replication forks occur by two different pathways. EMBO J. **16**:3332–3340.

Bierne, H., and B. Michel. 1994. When replication forks stop. Mol. Microbiol. **13**:17–23.

Brumbley, S. M., B. F. Carney, and T. P. Denny. 1993. Phenotype conversion in *Pseudomonas solanacearum* due to spontaneous inactivation of PhcA, a putative LysR transcriptional regulator. J. Bacteriol. **175**:5477–5487.

Bulmer, M. 1987. Coevolution of codon usage and transfer RNA abundance. Nature **325**:728–730.

Bussière, D. E., and D. Bastia. 1999. Termination of DNA replication of bacterial and plasmid chromosomes. Mol. Microbiol. **31**:1611–1618.

Capiaux, H., F. Cornet, J. Corre, M. I. Guijo, K. Perals, J. E. Rebollo, and J. M. Louarn. 2001. Polarization of the *Escherichia coli* chromosome. A view from the terminus. Biochimie **83**:161–170.

Cox, M. M., M. F. Goodman, K. N. Kreuzer, D. J. Sherratt, S. J. Sandler, and K. J. Marians. 2000. The importance of repairing stalled replication forks. Nature **404**:37–41.

Doolittle, W. F. 2000. The nature of the universal ancestor and the evolution of the proteome. Curr. Opin. Struct. Biol. **10**:355–358.

Eisen, J. A., J. F. Heidelberg, O. White, and S. L. Salzberg. 2000. Evidence for symmetric chromosomal inversions around the replication origin in bacteria. Genome Biol. **1**:research0011.1–0011.9.

Finkel S. E., and R. C. Johnson. 1992. The Fis protein: it's not just for DNA inversion anymore. Mol. Microbiol. **6**: 3257–3265.

Frank, C., and J. R. Lobry. 2000. Oriloc: prediction of replication boundaries in unnanotated bacterial chromosomes. Bioinformatics **16**:560–561.

Garcia-Vallvé, S., A. Romeu, and J. Palau. 2000. Horizontal gene transfer in bacterial and archaeal complete genomes. Genome Res. **10**:1719–1725.

Gonçalves, I., M. Robinson, G. Perrière, and D. Mouchiroud. 1999. JaDis: computing distances between nucleic acid sequences. Bioinformatics **15**:424–425.

Gouy, M., and C. Gautier. 1982. Codon usage in bacteria: correlation with gene expressivity. Nucleic Acids Res. **10**:7055–7073.

Gregg, A. V., P. McGlynn, R. P. Jaktaji, and R. G. Lloyd. 2002. Direct rescue of stalled DNA replication forks via the combined action of PriA and RecG helicase activities. Mol. Cell **9**:241–251.

Guindon, S., and G. Perrière. 2001. Intragenomic base content variation is a potential source of biases when searching for horizontally transferred genes. Mol. Biol. Evol. **18**: 1838–1840.

Hiasa, H., and K. J. Marians. 1992. Differential inhibition of the DNA translocation and DNA unwinding activities of DNA helicases by the *Escherichia coli* Tus protein. J. Biol. Chem. **267**:11379–11385.

Hill, T. M. 1992. Arrest of bacterial DNA replication. Annu. Rev. Microbiol. **46**:603–633.

Himmelreich, R., H. Plagens, H. Hilbert, B. Reiner, and R. Herrmann. 1997. Comparative analysis of the genomes of the bacteria *Mycoplasma pneumoniae* and *Mycoplasma genitalium*. Nucleic Acids Res. **25**:701–712.

Hudson, R. E., U. Bergthorsson, J. R. Roth, and H. Ochman. 2002. Effect of chromosome location on bacterial mutation rates. Mol. Biol. Evol. **19**:85–92.

Ide, H., H. Murayama, S. Sakamoto, K. Makino, K. Honda, H. Nakamuta, M. Sasaki, and N. Sugimoto. 1995. On the mechanism of preferential incorporation of dAMP at abasic sites in translesional DNA synthesis. Role of proofreading activity of DNA polymerase and thermodynamic characterization of model template-primers containing an abasic site. Nucleic Acids Res. **23**:123–129.

Ikemura, T. 1981. Correlation between the abundance of *Escherichia coli* transfer RNAs and the occurrence of the respective codons in its protein genes: a proposal for asynonymous codon choice that is optimal for the *E. coli* translational system. J. Mol. Biol. **151**:389–409.

Jensen, J. S., H. T. Hansen, and K. Lind. 1996. Isolation of *Mycoplasma genitalium* strains from the male urethra. J. Clin. Microbiol. **34**:286–291.

Kanaya, S., Y. Yamada, Y. Kudo, and T. Ikemura. 1999. Studies of codon usage and tRNA genes of 18 unicellular organisms and quantification of *Bacillus subtilis* tRNAs: gene expression level and species-specific diversity of codon usage based on multivariate analysis. Gene **238**:143–155.

Kaneko, T., Y. Nakamura, S. Sato et al. (21 co-authors). 2000. Complete genome structure of the nitrogen-fixing symbiotic bacterium *Mesorhizobium loti*. DNA Res. **7**:331–338.

Karlin, S. 1998. Global dinucleotide signatures and analysis of genomic heterogeneity. Curr. Opin. Microbiol. **1**:598–610.

Karlin, S., A. M. Campbell, and J. Mràzek. 1998. Comparative DNA analysis across diverse genomes. Annu. Rev. Genet. **32**:185–225.

Karlin, S., J. Mràzek, and A. M. Campbell. 1997. Compositional biases of bacterial genomes and evolutionary implications. J. Bacteriol. **179**:3899–3913.

Kerr, A. R., J. F. Peden, and P. M. Sharp. 1997. Systematic base composition variation around the genome of *Mycoplasma genitalium*, but not *Mycoplasma pneumoniae*. Mol. Microbiol. **25**:1177–1179.

Kunst, F., N. Ogasawara, I. Moszer et al. (148 co-authors). 1997. The complete genome sequence of the Gram-positive bacterium *Bacillus subtilis*. Nature **390**:249–256.

Lawrence, J. G., and H. Ochman. 1997. Amelioration of bacterial genomes: rates of change and exchange. J. Mol. Evol. **44**:383–397.

———. 1998. Molecular archaeology of the *Escherichia coli* genome. Proc. Natl. Acad. Sci. USA **95**:9413–9417.

Lemon, K. P., I. Kurtser, and A. D. Grossman. 2001. Effects of replication termination mutants on chromosome partitioning in *Bacillus subtilis*. Proc. Natl. Acad. Sci. USA **98**:212–217.

Lewis, P. J. 2001. Bacterial chromosome segregation. Microbiology **147**:519–526.

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

Lopez, P., H. Philippe, H. Myllykallio, and P. Forterre. 1999. Identification of putative chromosomal origins of replication in Archaea. Mol. Microbiol. **32**:883–886.

McGlynn, P., and R. G. Lloyd. 2001. Rescue of stalled replication forks by RecG: simultaneous translocation on the leading and lagging strand templates supports an active DNA unwinding model of fork reversal and Holliday junction formation. Proc. Natl. Acad. Sci. USA **98**:8227–8234.

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

McNeill, S. A. 2001. Understanding the enzymology of archaeal DNA replication: progress in form and function. Mol. Microbiol. **40**:520–529.

Médigue, C., T. Rouxel, P. Vigier, A. Hénaut, and A. Danchin. 1991. Evidence for horizontal gene transfer in *Escherichia coli* speciation. J. Mol. Biol. **222**:851–856.

Moran, N. A. 1996. Accelerated evolution and Muller's rachet in endosymbiotic bacteria. Proc. Natl. Acad. Sci. USA **93**:2873–2878.

Myllykallio, H., P. Lopez, P. Lopez-Garcia, R. Heilig, W. Saurin, Y. Zivanovic, H. Philippe, and P. Forterre. 2000. Bacterial mode of replication with eukaryotic-like machinery in a hyperthermophilic archaeon. Science **288**:2212–2215.

Ng, W. V., S. P. Kennedy, G. G. Mahairas et al. (40 co-authors). 2000. Genome sequence of Halobacterium species NRC-1. Proc. Natl. Acad. Sci. USA **97**:12176–12181.

Ochman, H., J. G. Lawrence, and E. A. Groisman. 2000. Lateral gene transfer and the nature of bacterial innovation. Nature **405**:299–304.

O'Reilly, M., and K. M. Devine. 1997. Expression of AbrB, a transition state regulator from *Bacillus subtilis*, is growth phase dependent in a manner resembling that of Fis, the nucleoid binding protein from *Escherichia coli*. J. Bacteriol. **179**:522–529.

Pedersen, A. G., L. J. Jensen, S. Brunak, H. H. Staerfeldt, and D. W. Ussery. 2000. A DNA structural atlas for *Escherichia coli*. J. Mol. Biol. **299**:907–930.

Perals, K., F. Cornet, Y. Merlet, I. Delon, and J. M. Louarn. 2000. Functional polarization of the *Escherichia coli* chromosome terminus: the *dif* site acts in chromosome dimer resolution only when located between long stretches of opposite polarity. Mol. Microbiol. **36**:33–43.

Perals, K., H. Capiaux, J. B. Vincourt, J. M. Louarn, D. J. Sherratt, and F. Cornet. 2001. Interplay between recombination, cell division and chromosome structure during chromosome dimer resolution in *Escherichia coli*. Mol. Microbiol. **39**:904–913.

Perrière, G., B. Labedan, and P. Bessières. 2000. EMGLib: the enhanced microbial genomes library (update 2000). Nucleic Acids Res. **28**:68–71.

Ragan, M.A. 2001. On surrogate methods for detecting lateral gene transfer. FEMS Microbiol. Lett. **201**:187–191.

Rocha, E. P. C., and A. Blanchard. 2002. Genomic repeats, genome plasticity and the dynamics of *Mycoplasma* evolution. Nucleic Acids Res. **30**:2031–2042.

Salanoubat, M., S. Genin, F. Artiguenave et al. (25 co-authors). 2001. Genome sequence of the plant pathogen *Ralstonia solanacearum*. Nature **415**:497–502.

Sawitzke, J., and S. Austin. 2001. An analysis of the factory model for chromosome replication and segregation in bacteria. Mol. Microbiol. **40**:786–794.

Schneider, R., R. Lurz, G. Luder, C. Tolksdorf, A. Travers, and G. Muskhelishvili. 2001. An architectural role of the *Escherichia coli* chromatin protein FIS in organising DNA. Nucleic Acids Res. **29**:5107–5114.

Sharp, P. M., and W. H. Li. 1987. The codon adaptation index—a measure of directional synonymous codon usage bias, and its potential applications. Nucleic Acids Res. **15**: 1281–1295.

Sharp, P. M., D. C. Shields, K. H. Wolfe, and W. H. Li. 1989. Chromosomal location and evolutionary rate variation in enterobacterial genes. Science **246**:808–810.

She, Q., R. K. Singh, F. Confalonieri et al. (28 co-authors). 2001. The complete genome of the crenarchaeon *Sulfolobus solfataricus* P2. Proc. Natl. Acad. Sci. USA **98**: 7835–7840.

Shields, D. C., and P. M. Sharp. 1987. Synonymous codon usage in *Bacillus subtilis* reflects both translational selection and mutation biases. Nucleic Acids Res. **15**:8023–8040.

Srivastava, A. K., and D. Schlessinger. 1990. Preparation of extracts and assay of ribosomal RNA maturation in *Escherichia coli*. Methods Enzymol. **181**:355–366.

Strauss, B. S. 1991. The 'A rule' of mutagen specificity: a consequence of DNA polymerase bypass of non-instructional lesions? Bioessays **13**:79–84.

Sueoka, N. 1962. On the genetic basis of variation and heterogeneity of DNA base composition. Proc. Natl. Acad. Sci. USA **48**:582–592.

———. 1992. Directional mutation pressure, selective constraints, and genetic equilibria. J. Mol. Evol. **34**:95–114.

Travers, A., R. Schneider, and G. Muskhelishvili. 2001. DNA supercoiling and transcription in *Escherichia coli*: The FIS connection. Biochimie **83**:213–217.

Tsai, L., and Z. Sun. 2001. Dynamic flexibility in the *Escherichia coli* genome. FEBS Lett. **507**:225–230.

Ussery, D., T. S. Larsen, K. T. Wilkes, C. Friis, P. Worning, A. Krogh, and S. Brunak. 2001. Genome organisation and chromatin structure in *Escherichia coli*. Biochimie **83**: 201–212.

Wake, R. G. 1997. Replication fork arrest and termination of chromosome replication in *Bacillus subtilis*. FEMS Microbiol. Lett. **153**:247–254.

Wang, B. 2001. Limitations of compositional approach to identifying horizontally transferred genes. J. Mol. Evol. **53**:244–250.

White, O., J. A. Eisen, J. F. Heidelberg et al. (32 co-authors). 1999. Genome sequence of the radioresistant bacterium *Deinococcus radiodurans* R1. Science **286**:1571–1577.

Yang, Z. 1997. PAML: a program package for phylogenetic analysis by maximum likelihood. Comput. Applic. Biosci. **13**:555–556.

William Martin, Associate Editor

Accepted September 8, 2002