The replication-related organization of bacterial genomes

Eduardo P. C. Rocha

Atelier de Bioinformatique, Université Pierre et Marie Curie, 12, Rue Cuvier, 75005 Paris, and Unité Génétique des Génomes Bactériens, Institut Pasteur, 28 rue du Dr Roux, 75724 Paris Cedex 15, France

The replication of the chromosome is among the most essential functions of the bacterial cell and influences many other cellular mechanisms, from gene expression to cell division. Yet the way it impacts on the bacterial chromosome was not fully acknowledged until the availability of complete genomes allowed one to look upon genomes as more than bags of genes. Chromosomal replication includes a set of asymmetric mechanisms, among which are a division in a lagging and a leading strand and a gradient between early and late replicating regions. These differences are the causes of many of the organizational features observed in bacterial genomes, in terms of both gene distribution and sequence composition along the chromosome. When asymmetries or gradients increase in some genomes, e.g. due to a different composition of the DNA polymerase or to a higher growth rate, so do the corresponding biases. As some of the features of the chromosome structure seem to be under strong selection, understanding such biases is important for the understanding of chromosome organization and adaptation. Inversely, understanding chromosome organization may shed further light on questions relating to replication and cell division. Ultimately, the understanding of the interplay between these different elements will allow a better understanding of bacterial genetics and evolution.

Background

The study of bacterial chromosome organization has received little attention, at least when compared with the focused research on the features of the eukaryotic chromosome. There are several historical reasons for this. First, bacterial chromosomes are much more homogeneous than most eukaryotic chromosomes in terms of sequence composition and gene density. Second, in the majority of bacteria there are no large characteristically diverse regions, such as isochores, telomeres and centromeres. Third, bacterial chromosomes vary between 500 kb and 10 Mb, which is significant, but much less than the length variation shown by eukaryotic chromosomes. Nonetheless, bacteria can have more than one chromosome, which can be linear or circular, or oscillate between the two (Casjens, 1998). There is also extensive size variation within a species, with strains of some species differing by more than 50% in genome size (Bergthorsson & Ochman, 1998; Carlson & Kolsto, 1994). This, together with the results coming from genome sequencing, sparked a new interest in this domain (reviewed by Casjens, 1998; Kolsto, 1997). The abundance of PFGE studies and especially the existence of more than 100 publicly available complete bacterial genomes have rendered possible the build-up of comparative genomic studies at very diverse levels of detail. A major result of such studies was the demonstration that replication is at the very basis of the organization of bacterial chromosomes. By this, I denote that chromosome replication is a major cause of the arrangement and interrelation of many of the elements constituting the chromosomal organization.

In this article, I review the more recent developments in the understanding of these trends and try to put them together in a unifying framework. I start with a brief outline of the mechanisms involved in bacterial replication, with emphasis on the asymmetries it leads to within and between bacterial genomes. I then discuss the major types of replication-associated asymmetries: differences between the leading and lagging strands and variations along the replicores in relation to the distance from the origin of replication. For each of these elements, biases at the level of nucleotides, oligonucleotides and genes are discussed and put into relation with current working hypotheses. This review is not about replication mechanisms, replication fidelity or the distribution of signals directly related to replication or its regulation. Although these elements also contribute to fashioning the bacterial chromosome, I highlight the implications of replication on the chromosome at wider organizational levels and its implications on mechanisms and structures other than replication itself. The emphasis is on the elements that are structured by...
replication without directly intervening in it. The argument is that the intrinsic mechanistic asymmetries or gradients of chromosome replication impose biases on sequence composition and gene distribution in the bacterial chromosome. For further information on the mechanisms and regulation of replication start, elongation and termination, the reader is invited to consult the excellent reviews available (Bussiere & Bastia, 1999; Giraldo, 2003; Kunkel & Bebenek, 2000; Lobry & Louarn, 2003; Marians, 1992).

**Replication in bacteria**

In bacteria where replication has been thoroughly studied, such as *Escherichia coli* and *Bacillus subtilis*, chromosome replication starts at a single chromosome locus, the origin (ori). The key regulation of chromosome replication takes place at the start of the process and is tightly coupled to cell mass (Boye et al., 1996). Initiation occurs once per cell cycle during a short time interval at all origins within a cell (Skarstad et al., 1986). Secondary initiations of the newly replicated origins are then avoided by a transient, temporally coordinated blockage in re-initiation at the origin (Campbell & Kleckner, 1990). Replication proceeds by the bidirectional progression of the replication forks along the chromosome (Fig. 1). In each replication fork, a complex with two DNA polymerases replicates the two DNA strands. The pace of the replication fork varies significantly between bacteria. For example, in *E. coli* the fork progresses at ~1000 nt s⁻¹, in *Pyrococcus abyssi* at ~300 nt s⁻¹ (Myllykallio et al., 2000) and in *Mycoplasma capricolum* at ~100 nt s⁻¹ (Seto & Miyata, 1998). Thus although the *M. capricolum* genome is seven times smaller than that of *E. coli*, it takes longer to be replicated. Short DNA segments to which a replication terminator protein binds (ter sites) arrest the movement of the replication forks when these pass through the terminus region (Bussiere & Bastia, 1999). The chromosome dimer is then resolved by site-specific recombination mediated by XerC and XerD recombinases at the *dif* sites (Blakely et al., 1993; Kuempel et al., 1991). The system of replication fork arrest diverges widely among bacteria, and deletion of ter sites is not lethal in *B. subtilis* (Iismaa & Wake, 1987) or in *E. coli* (Henson & Kuempel, 1985). On the contrary, the system of chromosome resolution is strongly conserved (Recchia & Sherratt, 1999).

**The replication fork**

Elongation of the newly synthesized strand involves the displacement of the replication fork along the chromosome. The replication fork is a holoenzyme with several components (Glover & McHenry, 2001), among which are a helicase that unwinds the chromosome (DnaB in *E. coli*), and two DNA polymerases III (DNAP) that replicate the two strands (Fig. 1). The DNAP core is composed of a heterodimer of the subunits α, ε and θ, and contains the polymerase activity (α-subunit) and the proofreading 3’→5’ exonuclease activity (θ-subunit). The θ-subunit causes DNAP to dimerize (Marians, 1992). Because DNA polymerization occurs in the 5’→3’ direction, one strand is replicated continuously – the leading strand – whereas the other strand is replicated in discrete steps – the lagging strand – through the use of the Okazaki fragments. The cycle of the synthesis of Okazaki fragments starts by the

![Fig. 1](https://example.com/fig1.png)

**Fig. 1.** Replication of the bacterial chromosome proceeds bidirectionally from the origin to the terminus of replication. At each replication fork, two DNAPs replicate the leading and lagging strands. Boxes indicate elements of asymmetry. See the text for details.
synthesis of a new 10–12 nt RNA primer by a primase (Kitani et al., 1985). Subsequently, the lagging strand polymerase transits from the 3′-OH terminus of the just completed Okazaki fragment to the new primer terminus, resuming DNA synthesis through 1000–2000 nt in E. coli and around 100 nt in P. abyssi (Matsunaga et al., 2002). The cycle ends by the removal of the previous RNA primer and gap filling. Conversely, the synthesis of the leading strand is essentially continuous, closely following the unwinding of the DNA duplex, with processivities surpassing 500 kb and eventually all the replichore (Marians, 1992).

Thus there are two polymerase cores within one holoenzyme particle, each replicating a different DNA strand. In E. coli, the two core polymerases are not pre-dedicated to one strand and can be interchanged without differences in processivity (Yuzhakov et al., 1996). In B. subtilis and the other Firmicutes there are two different genes encoding the α-subunit of the DNA polymerase (Bruck & O’Donnell, 2000; Dervyn et al., 2001). One of these subunits is the orthologue of the E. coli subunit (DnaE), whereas the other shows weak sequence similarity (PolC) (Koonin & Bork, 1996). Recent evidence suggests complementary roles for these proteins, with DnaE involved in lagging strand synthesis and PolC involved in leading strand synthesis (Dervyn et al., 2001). An important difference between the two proteins is that PolC includes the exonuclease domain, which is coded separately in E. coli (α-subunit, DnaQ) (Huang & Ito, 1999). A third family of α-subunits resembling E. coli’s DnaE is present in cyanobacteria (Kaneko et al., 1996). In this case the α-subunit is spliced in two genes (Ito et al., 1998). Finally, some bacteria show a more mixed, and less well-understood, picture. For example, in Thermotoga maritima, there is one gene encoding a homologue of PolC, but no obvious homologue of DnaE (Nelson et al., 1999).

Asymmetries and gradients in replication

The mechanistic asymmetries of bacterial replication result in mutational or selective biases, differentiating the chromosome within replichores and between replicating strands (Fig. 1).

Collisions between polymerases. Chromosomal replication often takes place in moments of high transcription levels and collisions between DNAP and RNA polymerase (RNAP) are inevitable. Although the RNAP transcription rate varies with the growth phase, it is usually in the range 40–50 nt s⁻¹, thus is 20 times slower than that of DNAP in E. coli (Bremer & Dennis, 1996). Because both polymerases progress in the 5′→3′ direction, transcripts coded in the lagging strand lead to head-on collisions whereas leading strand transcripts lead to co-oriented collisions (Fig. 2). There are two important consequences of this asymmetry. When the replication fork arrives at an actively transcribed operon, polymerases will inevitably collide if the operon is in the lagging strand but occur with probability ~19/20 (i.e. 95%) if the operon is in the leading strand (because DNAP is 20 times faster than RNAP). Thus the probability of collision depends on the direction of transcription. But so does the outcome of the collision. Studies of both E. coli rRNA (French, 1992) and Saccharomyces cerevisiae tRNA genes (Deshpande & Newlon, 1996) in vivo indicate that only head-on collisions interfere significantly with the progression of the replication fork. As a consequence, differential probability and outcome of collisions between polymerases lead to an asymmetric distribution of genes between the two strands.

Compositional differences between replicating strands. During the replication of the Okazaki fragments, the leading strand is kept single-stranded while the neo-formed lagging strand is being synthesized. Yet the lagging strand stays double-stranded when the neo-formed leading strand is being synthesized. Single-stranded DNA (ssDNA) can form secondary structures and these may lead to replication errors (Leach, 1994) and palindrome deletion (Trinh & Sinden, 1991). Also, some types of substitutions increase in ssDNA more than others, and thus the different replication of the two strands leads to compositional differences between the replicating strands (Francino & Ochman, 1997; Frank & Lobry, 1999).

Gene dosage effects. Fast-growing bacteria have growth rates requiring replication re-initiation before the round in progress is complete. In this way, E. coli can attain growth rates of 2·5 doublings h⁻¹ (Bremer & Dennis, 1996). The replication fork moves at about 600–1000 nt s⁻¹ (Marians, 1992), and to duplicate the average 5 Mb E. coli chromosome it takes from 40 to 67 min (Bremer & Dennis, 1996). Hence E. coli cells under exponential growth show two to three simultaneous rounds of replication, a new one starting every ~20 min. Under these conditions, genes near the origin of replication are overrepresented in the bacterial cell by a factor of 4 (2ⁿ) to 8 (2ⁿ) relative to genes near the terminus of replication (Chandler & Pritchard, 1975). Hence in fast-growing bacteria highly expressed genes tend to be positioned near the origin of replication.

Gradients along the replichore. The directionality of chromosome replication leads to a chronological asymmetry along the replichore. Since about 1 h separates the beginning from the end of replication, the favourable conditions leading to replication start may no longer prevail at the end of the process and this may result in different mutation rates and compositional biases (Daubin & Perriere, 2003). Dimer separation after full chromosome replication requires site-specific or homologous recombination (Corre & Louarn, 2002). This may also result in differential mutation and recombination biases near the terminus of replication.

In the following sections, these different biases are discussed in the framework of the major replication asymmetries. Additional material with the description of genomes used for the figures can be found in the online version of this.
Differences between leading and lagging strands

**Gene strand bias**

The ‘polymerase collision avoidance’ model. The seven rDNA operons in *E. coli* are all on the leading strand, resulting in co-orientation of replication and transcription (Ellwood & Nomura, 1982), which was speculated to result from selection to prevent too frequent collisions between DNAP and RNAP (Nomura & Morgan, 1977). These and other works showing that RNAP movement slows down DNA replication (Pato, 1975) led to the proposition that the differential outcome of collisions between DNAP and RNAP should result in a selective pressure towards coding highly expressed genes in the leading strand (Brewer, 1988). The lower collision rate would carry two major advantages: (i) faster DNA replication; (ii) less transcript loss. It could also diminish the number of replication arrests, which are dangerous for the cell (Kuzminov, 2001). In fast-growing *E. coli* there are ~70 000 ribosomes per cell, corresponding to roughly 10 000 transcripts per rRNA operon per replication cycle (Bremer & Dennis, 1996). Under the same conditions, there are about 70 RNAPs transcribing each of the seven rRNA operons, implicating a potential for about 500 collisions per replication round, i.e. about 5% of the rRNA transcription. This could significantly retard replication if collisions are head-on (Brewer, 1988). Indeed, highly expressed translation-related genes, such as rRNA and ribosomal protein genes, have been systematically found coded in the leading strands of genomes for which replicating strands can be identified (e.g. Blattner et al., 1997; Rocha, 2002; Zeigler & Dean, 1990). A rare exception concerns the sequenced genome of *Pasteurella multocida*, which exhibits two recent inversions near the origin of replication that shifted two of the six rDNA operons and several ribosomal protein genes to the lagging strand (May et al., 2001). The validity of the polymerase collision avoidance model has never been thoroughly tested, due to obvious experimental difficulties, and only recently has started to be questioned.

Frequency of genes in the leading strand. The sequencing of the regions surrounding replication origins of both *B. subtilis* (Ogasawara & Yoshikawa, 1992) and *E. coli* (Burland et al., 1993) showed that the genes are systematically coded in the leading strand. However, sequencing of
the two complete genomes resulted in very different observations (Fig. 3). The frequency of leading strand genes is \(~75\%\) in \textit{B. subtilis} (Kunst et al., 1997), but only \(~55\%\) in \textit{E. coli} (Blattner et al., 1997). A first systematic survey of gene strand bias showed that genomes could have from 55 to 80\% of genes in the leading strand, although systematically more than 85\% of ribosomal proteins were coded in the leading strand of these genomes (McLean et al., 1998). In fact, 78\% of the genes of Firmicutes (including mycoplasmas) are in the leading strand, compared with 58\% for the other genomes (Rocha, 2002). Interestingly, the group with higher biases coincides with the group of genomes containing two different (and probably strand-dedicated; Dervyn coincides with the group of genomes containing two \textit{Rocha, 2002}).

Interestingly, the frequency of leading strand genes is \(~55\%\) in \textit{Thermoanaerobacter tengcongensis} and \textit{Treponema pallidum}, and \textit{85\%} in \textit{B. subtilis}. The conjugation of these differences might render \textit{B. subtilis} DNAP more sensitive to collisions and thus favour higher gene strand bias. However, collisions between polymerases are expected to occur between RNAP and the helicase (Fig. 2), probably by DNA knotting and without direct contact (Olavarrieta et al., 2002). Although this does not necessarily invalidate the possibility that differential DNAP sensitivity could lead to higher overall biases, further experimental work will be necessary to assess this hypothesis.

Types of strand biased genes. Besides showing that compositionally different DNAPs are correlated with different levels of gene strand bias, these works raise another important point: high levels of gene expression can hardly account for all the observed trends. The number of genes that are highly expressed in bacteria is typically low (Andersson & Kurland, 1990), and cannot justify the high frequency of leading strand genes in Firmicutes. Furthermore, according to the polymerase collision model, gene strand bias should be higher in fast-growing bacteria, where transcription and replication (hence collisions) are very frequent. However, one typically finds the inverse (Rocha, 2002). That expression is not a determinant of gene strand bias was recently demonstrated in \textit{B. subtilis} and \textit{E. coli} (Rocha & Danchin, 2003a), where essentiality seems to be the unexpected key to understanding this bias. In \textit{B. subtilis}, the frequency of leading strand essential genes (96\%) and non-essential genes (74\%) is radically different and is independent of expression levels. Qualitatively similar results are found in \textit{E. coli} when high expression is defined using codon usage biases, or transcriptome or proteome data. Furthermore, these results

**Fig. 3.** Density of coding sequences (grey) and leading strand coding sequences (black) in model bacteria and highly biased genomes of Firmicutes (\textit{B. subtilis} and \textit{Thermoanaerobacter tengcongensis}) and others (\textit{E. coli} and \textit{Treponema pallidum}). The densities were computed in non-overlapping windows of 10 (\textit{T. tengcongensis} and \textit{T. pallidum}) and 20 (\textit{E. coli} and \textit{B. subtilis}) kb.
seem to hold when essentiality is assigned by homology in most other bacterial genomes (Rocha & Danchin, 2003b). In all cases, essential genes are more biased than non-essential genes, and among non-essential genes there is rarely a significant effect attributable to expression level. Finally, when comparing the location of orthologues in close genomes, essential genes are conserved in the leading strand more often than the other genes.

Re-thinking the polymerase collision model. The polymerase collision model does not satisfactorily explain the linkage between gene strand bias and essentiality. If the major problem associated with collision between polymerases is replication slow-down, then one would expect higher biases among the genes leading to higher collision rates, i.e. among highly expressed genes. Yet among the >3000 genes expressed during exponential growth in E. coli (Tao et al., 1999), the essential genes, many of which are expressed at low levels, are highly biased, whereas the highly expressed genes that are not essential are equally distributed between the replicating strands. This suggests that the problem of collisions is not related to the collision rate (i.e. with expression levels), but with gene function. In co-oriented collisions, the transcript may be finished, whereas head-on collisions result in aborted transcripts (Fig. 2). The latter may be translated into truncated non-functional peptides, which is particularly deleterious for essential functions. Also, non-functional peptides forming part of large complexes, as is often the case of essential genes, are typically dominant negative (Pakula & Sauer, 1989). As a result, the truncated peptide may inactivate an entire complex. If this model proves correct, it suggests that aborted transcripts may be more deleterious than previously thought.

It is yet unclear if high levels of gene strand bias in genomes containing two different DNAP α-subunits are related to a higher frequency of essential genes in the leading strand. One might consider that the problem of truncated peptides associated with collisions between polymerases is common to all expressed genes, but assumes a particular relevance in the context of genes encoding essential functions. If so, one could speculate that the asymmetry between the two types of collisions, head-on and co-oriented, would be qualitatively identical in all genomes, but have more severe consequences in genomes having two different DNAP α-subunits because they may be less robust to collisions, as discussed above. Hence the overall bias would be larger in these genomes.

**Compositional strand bias**

Early work suggested that the frequencies of G and C and A and T (\(n_G = n_C\) and \(n_A = n_T\)) are roughly equal in ssDNA (Rudner et al., 1968). These results were integrated in the theoretical body of molecular evolution, under the form of two parity rules (Sueoka, 1995). The first parity rule (PR1) indicates that when mutation and selection are symmetrical relative to both DNA strands, the rates of reciprocal substitutions are equal. The second parity rule indicates that at equilibrium, and under PR1 conditions, \(n_G = n_C\) and \(n_A = n_T\) in each DNA strand. In fact, under no strand bias the second parity rule is the expected outcome of the first (Lobry, 1995). However, the analysis of the first bacterial genomes showed that \(n_G = n_C\) and \(n_A = n_T\) do not hold when the genomes are divided into leading and lagging replicating strands (Lobry, 1996a). The subsequent flood of bacterial genomes confirmed the generality of this observation (Grigoriev, 1998; Lobry & Sueoka, 2002; Mackiewicz et al., 1999; McLean et al., 1998; Mrázek & Karlin, 1998; Rocha et al., 1999a). The analysis of the first 100 available complete genomes in the bacterial branch of the tree of life using conventional techniques indicates that less than a dozen lack any type of compositional strand bias (Fig. 4).

Several reviews have been dedicated to compositional strand bias (Francino & Ochman, 1997; Frank & Lobry, 1999; Karlin, 1999). Here I concentrate on the most recent results. With the exception of Streptomyces coelicolor (see below), the leading strand is richer in G relative to C, and to a lesser degree richer in T relative to A (Lobry, 1996a). Thus keto (leading strand) opposes amino (lagging strand) in compositional strand bias. The C-richness of the leading strands of mycoplasmas (McLean et al., 1998) is due to gene strand bias and biased composition of genes (Perrière et al., 1999). [Image: Fig. 4. Distribution of replication strand biases in the bacterial tree (tree adapted from Brochier et al., 2002).] Dark triangles indicate the ubiquitous presence of compositional strand bias, light triangles show groups containing some (usually few) genomes lacking compositional strand bias, and white indicates that all genomes sequenced so far seem to lack a significant bias. The hatched patterns distinguish the groups with particularly high levels of gene strand bias (on average more than 70% of leading strand genes). The black circle indicates the most commonly accepted position for the root of the tree.
et al., 1996), not compositional strand bias. Drawing GC skews, defined as (G−C)/(G+C) in sliding windows, has become a standard method to identify the origin and terminus of replication in many bacteria, where experimental evidence is rarely available (Grigoriev, 1998; Lobry, 1996b). Compositional strand bias is relatively homogeneous in the genome and is especially strong at third codon positions, as expected from a mutational bias (Lobry, 1996a; McLean et al., 1998; Rocha et al., 1999a). In highly biased genomes, strand bias is the major force shaping codon usage (McInerney, 1998; Romero et al., 2000). In these cases, one can easily discriminate the position of the genes relative to the replicating strand based on their codon usage or amino acid content of the respective proteins (Lafay et al., 1999; Mackiewicz et al., 1999; Rocha et al., 1999a). In Borrelia burgdorferi, one can predict with 95% accuracy the replicating strand positioning of the gene based solely on its amino acid content (97% using nucleotide composition). The accuracy of such discrimination analysis is also significant, although less pronounced, in the other genomes (Fig. 5). This is one of the most striking examples of a mutational bias shaping protein composition.

Strand bias is probably neutral and evolves fast. Genes that switch off replicating strands following chromosomal rearrangements evolve faster to adapt to the composition of the new replicating strand. As a result they show lower mean sequence similarity (Tillier & Collins, 2000b). The ratio of synonymous/non-synonymous rates of switched genes is similar to that of the average orthologues, in agreement with a mutational driving force for this process (Rocha & Danchin, 2001). As a consequence, strand biases and fast sequence adaptation after strand switch can lead to problems in estimating reliable phylogenies based on sequence analysis (Szczechapian et al., 2001).

Mechanism(s) causing compositional strand bias. Hypotheses aiming at identifying the causes of compositional strand bias must take into account the compositional changes it creates, its apparently neutral basis, and its near ubiquity.

Asymmetrical functioning and processivity of the two DNAPs at each replication fork could result in different mutation frequencies in the two strands, although there are conflicting reports on this issue (see Frank & Lobry, 1999, for a discussion on this subject). The lagging strand has been proposed to be less mutagenic based on the analysis of mutations in lacZ (Fijalkowska et al., 1998). The leading strand DNAP is highly processive staying on the template throughout replication, while the lagging complex needs to allow rapid cycling and may thus dissociate more easily from the template, leaving a mismatch for correction. However, lacZ is a native lagging strand gene, and adaptation to the leading strand composition is expected to involve higher substitution rates (Szczechapian et al., 2001). Radman (1998) proposed that the MutSHL mismatch repair system would be more efficient in correcting errors in the lagging strand because it requires DNA nicks to proceed and these are more readily available in the lagging strand. However, some genomes without the mismatch repair system have strand biases (e.g. Actinobacteria), and some genomes with MutSHL do not (e.g. Synechocystis and Anabaena). More importantly, all the above hypotheses allow an explanation of the different mutation rates between strands. However, the relevant variable regarding compositional strand bias is the substitutions’ asymmetry and different mutation rates per se do not lead to compositional asymmetry (Lobry & Lobry, 1999). Recent works identified compositional strand bias in archaea (Lopez et al., 1999), mitochondria (Mohr et al., 1999; Reyes et al., 1998) and phages (Grigoriev, 1999; Miller et al., 2003; Mrázek & Karlin, 1998). These elements possess different DNAP and repair mechanisms, different G+C compositions (Sueoka, 1962) and different mutation rates (Drake et al., 1998). Nevertheless, they all have a qualitatively similar compositional strand bias. Thus the major source of compositional asymmetry is probably to be found in a fundamental property such as the chemical stability of DNA.

The cytosine deamination theory proposes that compositional strand bias is caused by the chemical instability of cytosine in ssDNA (Frank & Lobry, 1999). Relative to dsDNA, the rate of cytosine deamination increases by a factor of 140 in ssDNA, and by a further factor of 4 when cytosine is methylated (see Lutsenko & Bhagwat, 1999, for a recent review). In the latter case, the deamination produces a T (instead of an U), which is not corrected by the uracil-DNA glycosylase (Coulondre et al., 1978). The leading strand is more exposed in the single-stranded state (in order to serve as template for the synthesis of the new
lagging strand), and C→T mutations would lead to GC and TA skews (and larger GC skews in G+C-poor genomes). Still, GC skews are larger than expected, relative to TA skews, under the cytosine deamination hypothesis (McLean et al., 1998). Data on inverted genes of *Chlamydia* and *Bacillus* were concordant with a major effect of cytosine deamination in establishing the bias, but only when a smaller contribution of asymmetric C→G substitutions was considered (Rocha & Danchin, 2001). Caution is necessary when interpreting the data of the latter analysis because the substitutions could not be oriented, and the genomes had accumulated multiple substitutions. Yet, when the joint contributions of cytosine deamination and C→G asymmetry were taken into consideration, there was a satisfactory explanation of the association between GC and TA skews in the available set of bacterial genomes (Rocha & Danchin, 2001). As a result, cytosine deamination of leading strand ssDNA is currently seen as the most likely and important cause of compositional strand bias. However, it may fail to describe the entire complexity of this phenomenon since *Streptomyces coelicolor* shows no significant compositional strand bias (Omura et al., 1998). It is still unknown if the inverted bias of *S. coelicolor* is indeed related directly to composition strand bias, or to an extreme GC composition (72%) or to the frequent inversions, fusions, deletions and insertions of large plasmids at these chromosomal regions in *Streptomyces* (Chen et al., 2002).

Variation in the intensity of compositional strand bias. As is clear from Fig. 5, compositional strand bias varies significantly from genome to genome. This may be related to the different stability of the genomes, as chromosome shuffling will tend to level off the bias (Achaz et al., 2003; Mackiewicz et al., 2001b; Rocha et al., 1999b; Tillier & Collins, 2000b). The different length of the Okazaki fragments may also contribute to modulating the bias, since a smaller exposure in the ssDNA state would lead to less cytosine deamination (Mrázek & Karlin, 1998). Consistent with this hypothesis, the intensity of compositional strand bias is positively correlated with the duration of the single-stranded state of the H-genes during mitochondrial replication (Reyes et al., 1998). Eukaryotes have Okazaki fragments that are 10 times smaller than those in *E. coli*, and lack extensive replication strand bias (Gierlik et al., 2000). *P. abyssi* and *Sulfolobus acidocaldarius* also have small Okazaki fragments (Matsumaga et al., 2002), which could explain the lower biases typically found in archaea (Lopez & Philippe, 2001). Finally, some bacteria are more exposed than others to DNA mutagenic agents. Obligately intracellular bacteria may be particularly well protected from this point of view, and this could partly explain why they tend to exhibit higher biases.

### Oligonucleotide bias

Oligonucleotide frequency and distribution in chromosomes is far from random, even when the relative frequencies of the sub-sequences they contain are taken into account (Karlin & Brendel, 1992). Further, oligonucleotide usage differs between the two replicating strands (Lopez et al., 1999; Mrázek & Karlin, 1998; Rocha et al., 1998; Salzberg et al., 1998). This is the result of the overlapping of different factors such as composition strand bias, replication signals or the asymmetric distribution of genes and corresponding regulatory signals in the two replicating strands. If a genome has a high translation-associated codon usage bias, the differential distribution of genes in the two strands may relate in a complicated way to composition strand bias (Musto et al., 2003; Romero et al., 2000). Therefore, oligonucleotide bias is the result of conflicting biases associated with signals related to different cellular processes (Trifonov, 1989). For example, the sequence signalling the ribosome-binding site (RBS) is very often found in the leading strand of *B. subtilis*. This is partly because it is G-rich and C-poor, but most fundamentally because 75% of the genes of this genome are in the leading strand. Similarly, stretches of T’s are often related to the presence of rho-independent terminators (d’Aubenton Carafa et al., 1990), and thus are more often found in the leading strand.

Chi sequences are *cis*-acting DNA elements that enhance recombination promoted by the RecBCD pathway in *E. coli* (Smith, 1988). These sequences are overrepresented (75%) in the leading strand of *E. coli* (Blattner et al., 1997), which allows collapsed or reversed replication forks to be protected against exonuclease degradation and to be preferentially used as a recombination substrate (Gruss & Michel, 2001). However, these sequences are GT-rich and include triplets corresponding to frequent codons. Thus doubts remain as to the biological significance of this overrepresentation in *E. coli* (Bell et al., 1998; Blaudet et al., 1998; Uno et al., 2000). Furthermore, very different chi sequences have been found in *Haemophilus influenzae* and *B. subtilis*, where the leading strand bias is much less pronounced (El Karoui et al., 1999). Such an overlapping of signals and biases complicates the understanding of the biological reasons behind most oligonucleotide avoidance or over-representation (Burge et al., 1992; Mrázek & Karlin, 1998; Salzberg et al., 1998).

### Gradients along the replicohores

#### Gene distribution

Gene dosage effects. When the time required to replicate the chromosome exceeds the duplication time, the dosage of genes near the origin in the cell increases exponentially with the number of simultaneous replication rounds. Fast-growing bacteria, such as *E. coli* or *B. subtilis*, with multiple simultaneous replication rounds selectively accumulate highly expressed genes near the replication origin because of this effect (Fig. 6). The moderately
fast-growing *Caulobacter crescentus* has an optimal duplication time of 90 min, which is more than that necessary to replicate its chromosome (i.e. ori/ter ≤ 2). Presumably this does not carry sufficient selective advantage to constrain these genes to the vicinity of the origin, and in slow-growing bacteria there is no appreciable bias in the distribution of highly expressed genes along the replicore (Fig. 6). Although gene dosage effects have rarely been tested experimentally, they have been frequently invoked to explain the deleterious effect of chromosomal rearrangements (Liu & Sanderson, 1996; Roth et al., 1996). Experimentally induced inversions in the *E. coli* chromosome altering the distance of genes to the origin of replication can lead to halving growth rates (Louarn et al., 1985). Similarly, systematic translocations of an expressed gene in the *Salmonella enterica* serovar Typhimurium chromosome indicate that its positioning closer to the origin leads to higher expression levels (Schmid & Roth, 1987). Essential genes are also not randomly placed along the replicore, but this is not an intrinsic feature of essential genes, being caused by the overrepresentation of highly expressed genes in this subset (Fig. 7). Thus among the constraints imposed by replication on the distribution of genes in the bacterial chromosome, expression plays a role in the distribution of genes as a function of the distance to the origin of replication, whereas essentiality constrains the replicating strand where genes are coded.

**rRNA operons.** In *E. coli* cells growing at exponential rate, rRNA corresponds to 80% of the total RNA and 50% of all transcription taking place in the cell (Bremer & Dennis, 1996). Although in slow-growing bacteria this proportion is likely to be smaller, rRNA components are arguably among the most highly expressed genes in bacteria. As a result, many genomes show a strong concentration of multiple rRNA operons near the origin of replication (Fig. 6). The exceptions to this trend are slow-growing bacteria such as *Mycoplasma* and *Mycobacterium*, which often have one single rRNA operon closer to the terminus than to the origin of replication. The compartmentalization of transcription and translation studied through GFP fusions in *B. subtilis* shows that Rnap is located within the nucleoid, whereas the ribosomes are located predominantly towards the cell poles close to or in contact with the membrane (Lewis et al., 2000). At higher growth rates transcription foci are close to the origin of replication, presumably because most highly expressed genes are located there. In *B. subtilis* the 10 rRNA operons are arranged in four groups (Kunst et al., 1997). The largest one contains seven operons placed very close to the origin of replication (< 175 kb), and recent data indicate that only these operons are included in transcription foci, even though the other operons are transcribed at significant rates (Davies & Lewis, 2003). Because of gene dosage effects, the operons most distant from ori account for less than 10% of all rRNA transcription at high growth rates in *B. subtilis*. Significant expression and regulatory response heterogeneity is also found among the seven *E. coli* operons, with operons further from the origin showing less expressiveness and/or smaller response to standard transcriptional activators (Condon et al., 1993). It is tempting, then, to speculate that ribosomes closer to the origin are dedicated to translating standard structural genes, such as ribosomal proteins, whereas the other ribosomes are dedicated to other types of transcripts. It has often been suggested that phenotypic heterogeneity in the translation machinery of bacteria could be selected to increase population adaptability (Mikkola & Kurland, 1991; Norris & Madsen, 1995). In this case, a determinant variable for the
assignment of ribosomes to their dedicated role would be the distance of the respective operons to the origin of replication.

**Phage integration and gene transfer**

The genomes of *B. subtilis* (Kunst *et al.*, 1997) and *E. coli* K12 (Blattner *et al.*, 1997) show A+T-rich *ter* regions associated with the presence of prophages. Since recombination is involved in the resolution of the replicated chromosomes and in the integration of horizontally transferred genes, it was tempting to relate the two. Horizontal gene transfer (HGT) was suggested to cluster closer to the terminus of replication (Campbell, 1992), and it has recently been suggested that hyper-recombination in the *E. coli ter* region is mostly caused by these sites (Corre *et al.*, 2000). The availability of a large number of genomes from bacteria closely related to *E. coli*, *B. subtilis* and *Streptococcus pneumoniae* allows the reappraisal of these analyses by identifying genes arising from HGT. The results suggest a very small bias towards larger rates of HGT in the second half of the chromosome for enterobacteria and *Streptococcus*, and smaller than expected HGT in the terminus of *Bacillus* (Fig. 8). Although the regions immediately adjacent to the origin seem to underrepresent HGT systematically, there is no clear bias for an overrepresentation of HGT at the terminus.

**Heterogeneities in nucleotide composition**

Early analyses of nucleotide composition in bacterial genomes showed high inter-genomic (Sueoka, 1962) but low intra-genomic (Rolfe & Meselson, 1958) variability. Further, each bacterial lineage has a characteristic pattern of codon usage, which is determined by several factors such as its G+C content, its tRNA set and gene expression levels (Grantham *et al.*, 1980; Ikemura, 1981). These observations led to the development of HGT detection methods based on sequence composition (Lawrence & Ochman, 1997; Médigue *et al.*, 1991). These methods assume that nucleotide composition is relatively uniform along the chromosome. However, G+C content among genes expressed at low levels is lower in late-replicating regions of the *E. coli* genome (Deschavanne & Filipski, 1995). It also shows significant variations, although not in function of replicore position, in *Mycoplasma genitalium* (Kerr *et al.*, 1997). This may lead to an overestimation of HGT by methods based on nucleotide composition, although real HGTs are most often found by these methods (Daubin *et al.*, 2003). These results have recently been confirmed and extended by an analysis using several complete genomes (Daubin & Perrier, 2003). Half of the sequenced genomes show A+T-richness at the terminus of replication, with a significant proportion of the others showing significant heterogeneities along the chromosome not coincident with the replication terminus. Tests done in *E. coli*, *Chlamydia* and *Helicobacter* showed that HGT was not the cause of these heterogeneities and one must then conclude that the composition of bacterial chromosomes is more heterogeneous than previously thought. Most of the heterogeneity is correlated with the organization of the chromosome relative to replication, although this leads to smaller differences in nucleotide composition than positional strand bias (Fig. 9).

The mechanisms establishing these replicore compositional
gradients are probably highly variable. The genome of Corynebacterium diptheriae shows a remarkably lower G+C content at the terminus, but the closely related, and largely co-linear, genome of Corynebacterium efficiens does not show such a pattern (Cerdon-Tarraga et al., 2003). The mutation rates of TA→GC transitions and transversions in a lacZ revertant in four loci along the chromosome of S. enterica serovar Typhimurium showed significant heterogeneities, but not a clear gradient in the direction origin to terminus (Hudson et al., 2002). However, AT-richness near the terminus is the result of the balance between GC→AT and TA→GC mutations, and this balance has yet to be analysed. Two other possible causes for replichore compositional gradients can be proposed. First, the recombination at dif sites may induce a bias associated with replication and/or repair of the end of both concatemates (see Daubin & Perriere, 2003, and references therein). Second, if the end of replication is taking place in conditions of depleted nutrients, then A+T-richness at the terminus could result from the relatively larger availability of these nucleotides in bacterial cells (Rocha & Danchin, 2002; Sharp et al., 1989). Exponentially growing bacteria are simultaneously replicating regions near the terminus and origin of replication, and no base composition heterogeneity should be observed under these circumstances. However, under low growth rates, for which the condition of depleted nutrients could be observed under these circumstances. However, under low growth rates, for which the condition of depleted nutrients would lead to an A and T enrichment at the terminus region. Since A+T-richness shows a significant increase just at the vicinity of the terminus, this would suggest a sudden decrease of the G+C nucleotide pool at the end of chromosome replication.

**Rate of sequence change**

Sharp et al. (1989) found that genes near the origin of replication of E. coli had synonymous substitution rates, after accounting for codon usage bias, that were a half smaller than those of genes near the terminus. This analysis included the available 67 pairs of orthologues between E. coli and S. enterica serovar Typhimurium, and the major effect was observed in genes very close to the origin (which contain many highly conserved genes). Recent analysis of closely
related complete genomes, including *E. coli* and *S. enterica* serovar Typhimurium (Mira & Ochman, 2002), confirmed the widespread existence of such bias. However, after controlling for expression levels, using measures of codon usage bias, the distance of the gene to the origin of replication accounts for only ~5% of the variance in substitution rates along the replicohore. More recent work has shown that synonymous substitution rates are relatively constant throughout the replicohore, with the exception of the G+C-poor regions immediately next to the terminus of replication (Daubin & Perriere, 2003). Thus it is tempting to correlate the variation of substitution rates with distance to the origin with the nucleotide bias along the replicohore that was mentioned in the previous section.

**Open questions for research**

**The coupling of replication biases and segregation**

Bacteria segregate their chromosomes without a mitotic apparatus. Following replication start, the origins rapidly move apart to opposite poles of the cell, whereas the terminus of replication is located at the cell centre. The remaining chromosome is between the pole and the cell centre (for a recent review see Errington et al., 2003). Recent work has suggested an important role of the chromosome replication structure in chromosome segregation, both at the origin and at the terminus of replication.

The intriguing rapid separation of replication origins after replication led to the search for sequence signals that might drive the process. Lin & Grossman (1998) identified a degenerate 16 bp DNA signal, present 10 times in the chromosome, always near the origin, eight instances of which were found to bind to the partitioning protein Spo0J *in vivo*. However, recent data show that these sites are not necessary for segregation (Wu & Errington, 2002), and the quest for such putative motifs continues. A very different view of the partition process suggests that the random diffusion of DNA is transiently constrained by the process of co-transcriptional translation and translocation (named transertion) of membrane proteins (Norris & Madsen, 1995; Woldringh et al., 1995). After initiation of DNA replication, the positioning of transertion areas near the origin of replication might create a bidirectional expansion force leading to the separation of the chromosomes (Woldringh, 2002). Recent evidence from genome analysis suggests that highly expressed genes placed asymmetrically relative to the origin of replication could also result in a mechanism of chromosome segregation. This could result from two different effects. First, the motor force of RNAP, whose movements are restricted in the cell, could pull the replication origins apart when transcribing the highly expressed genes near the origin (Dworkin & Losick, 2002). Second, highly expressed genes are distributed asymmetrically around the origin in both *B. subtilis* and *E. coli*, which coupled with transcriptional effects and hemi-methylated DNA could result in chromosome segregation and bacterial differentiation (Rocha et al., 2003).

The terminus region, being the last to be replicated and segregated, is present at the cell centre during termination (Errington et al., 2003). Site-specific recombination for chromosome dimer resolution is coupled to cell division by the presence of the FtsK protein at the *dif* site (Perals et al., 2001). FtsK forms a ring-like structure at the cell centre, and may be involved in guiding the regions flanking the *dif* sites in dimeric chromosomes towards the septum, so that synopsis and recombination can occur (Corre & Louarn, 2002; Li et al., 2003). A model has been proposed to explain the connection between segregation and dimer resolution (Corre & Louarn, 2002). When the septum is closing, the FtsK rings around the DNA threads mobilize them to place the *dif* site exactly under the septum. The recognition of the direction in which DNA should be pumped by the translocase would be made by the asymmetric distribution of polar motifs in the chromosome (Capiaux et al., 2001). Recent results suggest that such motifs could be widespread in bacterial genomes, thus connecting the opposing composition of leading and lagging strands with signals that are increasingly biased between the strands near the terminus of replication (Lobry & Louarn, 2003). The regular polarized distribution of these motifs could impose significant limitations to chromosome inversions and HGT (Lawrence & Hendrickson, 2003). All this suggests a much closer integration between replication, segregation, transcription and the chromosome structure than previously thought.

**Understanding replication from the organization it induces**

The analysis of replication biases in genomes has been useful in the understanding of replication itself. A first obvious use of this information resulted in the identification of putative replication origins in most bacterial genomes. Composition strand bias also directed experimental work to the identification of replication origins in the linear chromosome of *B. burgdorferi* (Picardeau et al., 1999) and the first identification of an origin of replication in an archaeon (Myllykallio et al., 2000). Yet some genomes lack all types of replication bias. These include several cyanobacteria, *Aquifex* and *Thermotoga* (Fig. 4) as well as most archaea. Could this be caused by different replication mechanisms, e.g. lack of a unique origin of replication (Mrázek & Karlin, 1998)? Interestingly, two recently sequenced genomes of closely related γ-proteobacteria obligate symbionts lack a gene encoding DnaA, opening the possibility that they might lack a unique origin of replication. However, whereas *Wigglesworthia glossinidia* lacks significant compositional strand bias (Akman et al., 2002), *Blochmannia floridanus* shows a very strong one (Gil et al., 2003). Therefore, although *W. glossinidia* is the only case of absence of significant compositional strand bias among γ-proteobacteria, this does not allow any conclusions to be drawn regarding the association of the loss of DnaA and strand biases. Recent experimental work suggests that *Methanocaldoccoccus jannaschii* may have more
than one replication origin (Maisnier-Patin et al., 2002). A genome with multiple potential replication origins would lack strand biases because successive replication rounds starting at different locations would be continuously redefining the replication strand of each region. This would level off any eventual biases.

Still, one wonders how genomes lacking replication-associated biases can overcome the problems caused by DNAP and RNAP collisions. Although many phages present strand biases (Grigoriev, 1999), other phages, such as T4 and φ29, have mechanisms to deal with head-on collisions. In T4 the absence of the τ-subunit of DNAP may prevent the physical coupling of the two DNAPs in the replicating fork (Alberts et al., 1983). Possibly as a result of this, it has been found that T4 replication forks can bypass oppositely oriented stalled transcription complexes in vitro (Liu & Alberts, 1995). The linear B. subtilis phage φ29 is replicated from both extremities in such a way that it does not have lagging strands and both transcription orientations result in head-on and co-oriented collisions. In this phage co-oriented collisions lead, as in E. coli, to replication slowdown, which is probably caused by the DNAP following an actively transcribing RNAP (Elias-Arnanz & Salas, 1997). On the other hand, head-on collisions are dealt with in such a way that the DNAP by-passes RNAP without displacing it, in a mechanism that is probably different from that of T4 (Elias-Arnanz & Salas, 1999). It remains to be investigated whether bacteria lacking replication bias have also found ways of dealing with the problem of collision between polymerases.

Challenges to the chromosomal structure

The availability of more than 100 bacterial genomes allows the comparative evolutionary analysis of genome structure and bacterial lifestyle. The most significant compositional strand biases have been found in obligately intracellular bacteria (Fig. 5). The reasons for this are not quite clear. Although these genomes have lost a significant part of their repair mechanisms, so have the small genomes of mycoplasmas, which do not have strong compositional biases but do have strong gene strand bias. The association between obligately intracellular niches and compositional strand bias may be caused by the extreme stability of most of these genomes (Tamas et al., 2002). Mycoplasmas, which are less stable, show lower compositional strand bias. Repeats induce frequent chromosomal rearrangements, and may thus reduce strand biases (Rocha et al., 1999b). Genomes depleted of repeats would then be more stable and thus accumulate larger compositional strand bias (Achaz et al., 2003; Frank et al., 2002). For example, B. burgdorferi has the strongest compositional strand bias and is nearly devoid of repeats. However, its plasmids contain a significant number of repeats, and a much smaller compositional strand bias (Picardieu et al., 2000). GC skews have been used to infer patterns of rearrangements between chromosomes (Grigoriev, 2000; Zivanovic et al., 2002), although the rapidity of strand adaptation limits this analysis to close genomes. Gene strand biases are relatively insensitive to different bacterial lifestyles (Rocha, 2002), but as shown above, the distribution of highly expressed genes along the replichore is highly biased for fast-growing bacteria.

These biases impose constraints on chromosomal rearrangements. On the other hand, rearrangement rates accelerate with the number of repeated elements in bacterial chromosomes (Rocha, 2003), which are maintained by either selfish systems (e.g. insertion sequences) or selection for elements capable of generating genetic diversity (e.g. for antigenic variation). It has been observed that repeats tend to be placed in the chromosome in such a way that the rearrangements they may induce provoke smaller disruptions to the chromosomal replication structure than expected by chance (Achaz et al., 2003). Further, there is an important negative correlation in bacteria between the number of repeats and the replication strand bias. This suggests a trade-off between the advantages associated with chromosomes organized in function of replication and the rearrangement events caused by repeats. Rearrangements involving the change of the positioning of genes relative to the origin of replication and from one replicating strand to another lead to lower growth rates (Campo et al., 2004; Louarn et al., 1985; Segall et al., 1988; Wu & Errington, 2002). The fitness impact of such rearrangements is correlated with the degree of replication-associated disorganization they induce in the chromosomal structure. In some cases, inversions seem to be more deleterious than sequence deletion. For example, E. coli strains show moderate fitness loss upon loss of the region around the terminus (Henson & Kuempel, 1985). However, many inversions near the terminus of this chromosome are non-permissive, allegedly because they disrupt the structure allowing correct termination and segregation of the chromosomes (Guijo et al., 2001). Given the organization of genomes relative to replication, rearrangements strongly disrupting the chromosome structure are naturally bound to be deleterious (Mackiewicz et al., 2001a). Therefore, most fixed large chromosomal rearrangements have led to small structural changes in the chromosome relative to the replication (Eisen et al., 2000; Suyama & Bork, 2001; Tillier & Collins, 2000c). The understanding of the interplay between organizing features of the chromosome, such as replication, and elements inducing sequence variation and chromosomal rearrangements will allow an explanation of why different genomes show such different levels of organization and how this relates with their evolutionary history and ecology.

Conclusion

Bacterial genomes do not show the heterogeneity in coding density and nucleotide composition characteristic of the genomes of many eukaryotes. Still, they do show important variations regarding several aspects of genome organization, many of which relate to the asymmetries induced by the mechanism of chromosome replication. Interestingly, one single key asymmetry may lead to different biases. For
example, compositional strand bias and gene strand bias both originate in the asymmetric nature of the replication fork that divides the chromosome in leading and lagging strands. Yet the two biases are poorly correlated (Rocha, 2002; Tillier & Collins, 2000a). This is because their direct cause is different: in one case asymmetric exposure of ssDNA creates asymmetric composition, whereas in the other case the collision between polymerases favours leading strand genes. In Table 1, I summarize the different types of bias that have been found and their mechanistic basis. I also tentatively assign to each bias a mutational or selective basis. In some cases, e.g. compositional strand bias or gene strand bias, such assignment is relatively straightforward, but in others it is not. It is also not unlikely that in some cases mutual biases have been recruited for other (eventually selective) purposes. In any case, comparative genome analysis will have to take such biases progressively into account in oligonucleotide frequency or even phylogenetic studies, since the evolution of the sequences clearly depends on their location in the chromosome. The current flood of bacterial genome data will probably provide new information relating the subjects discussed in this article. However, experimental work in organisms other than E. coli will be of utmost importance in understanding the diversity of biases associated with bacterial replication. The wealth of expression data coming from genomics and microbial cell biology may also allow a better understanding of the associations between replication associated biases and other structuralizing variables such as cell division, gene expression and compartmentalization of the cell.

Table 1. Comparative analysis of the different biases arising in chromosomes from the asymmetric replication mechanisms operating in bacterial cells

<table>
<thead>
<tr>
<th>Replication bias</th>
<th>Selection/mutation</th>
<th>Effect</th>
<th>Major basis of the bias</th>
</tr>
</thead>
<tbody>
<tr>
<td>Between strands</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Nucleotide composition</td>
<td>M</td>
<td>$G_{\text{lead}}/G_{\text{lag}} &gt; G_{\text{lead}}/G_{\text{lag}}$ and $T_{\text{lead}}/T_{\text{lag}} &gt; A_{\text{lead}}/A_{\text{lag}}$</td>
<td>Chemical vulnerability of ssDNA</td>
</tr>
<tr>
<td>Gene distribution I</td>
<td>S</td>
<td>Most essential genes are coded in the leading strand</td>
<td>Collisions between DNAP and RNAP</td>
</tr>
<tr>
<td>Gene distribution II</td>
<td>S</td>
<td>Most genes are coded in the leading strand</td>
<td>Collisions between DNAP and RNAP; DNAP composition</td>
</tr>
<tr>
<td>Palindrome deletion</td>
<td>M</td>
<td>Large palindromes are deleted</td>
<td>Slipped mispair in ssDNA</td>
</tr>
<tr>
<td>$\chi$ sites</td>
<td>S (?)</td>
<td>$\chi$ sites are more abundant in the leading strand</td>
<td>Recombination repair of stalled replication forks</td>
</tr>
<tr>
<td>Along replichores</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Gene distribution</td>
<td>S</td>
<td>Highly expressed genes cluster near the origin</td>
<td>Gene dosage in fast-growing bacterial cells</td>
</tr>
<tr>
<td>Nucleotide composition</td>
<td>M (?)</td>
<td>$A+T$-rich terminus</td>
<td>Phosphorylation, recombination (?), nucleotide scarcity (?)</td>
</tr>
<tr>
<td>Rate of sequence evolution</td>
<td>M (?)</td>
<td>Sequence divergence increases along the replichore</td>
<td>Recombination (?), nucleotide scarcity (?)</td>
</tr>
<tr>
<td>Both</td>
<td>S (?)</td>
<td>Some motifs are overrepresented in the leading strand and near $\text{dif}$ sites</td>
<td>Chromosome segregation and resolution (?)</td>
</tr>
</tbody>
</table>
Acknowledgements

I am much indebted to the persons with whom I’ve worked and discussed the issues related to this review, and in particular to Antoine Danchin, Alain Viari and Vic Norris. Antoine Danchin, Guillaume Achaz and Vincent Daubin have given helpful criticisms and suggestions on earlier versions of this manuscript. Irina Artsimovitch provided helpful comments on the differences between RNAP in B. subtilis and E. coli.

References


Achaz and Vincent Daubin have given helpful criticisms and sug-

References


Achaz and Vincent Daubin have given helpful criticisms and sug-

References


Achaz and Vincent Daubin have given helpful criticisms and sug-

References


Achaz and Vincent Daubin have given helpful criticisms and sug-


