ACADEMIC PRESS

Available online at www.sciencedirect.com

SCIENCE DIRECT®

Biochemical and Biophysical Research Communications 299 (2002) 126–134

BBRC

www.academicpress.com

Analysis of bilateral inverse symmetry in whole bacterial chromosomes

J. Sánchez[a,b,*] and M.V. José[c]

[a] Department of Medical Microbiology and Immunology, University of Gothenburg, Gothenburg SE413-46, Sweden
[b] Facultad de Medicina, UAEM, Av. Universidad 1001, Cuernavaca, Morelos CP62210, Mexico
[c] Instituto de Investigaciones Biomédicas, UNAM, Ciudad Universitaria, México, D.F. CP70228, Mexico

Received 15 October 2002

Abstract

The positions of the 64 DNA tri-nucleotides (triplets) along the *Borrelia burgdorferi* chromosome were determined and cumulative position plots (CPP) were obtained. Analysis of CPP for complementary triplets revealed close correlations in complementary triplet frequencies (CTF) between opposing leading and lagging strands. Such bilateral inverse symmetry (BIS) applied also to complementary mono- and di-nucleotides and to some >3 *n*-tuples. At the level of individual bases BIS explains Chargaff's second parity rule for whole bacterial chromosomes. Using shuffled control sequences we show that single-base BIS was not the source of higher-order BIS. Analysis of CTF in 45 other chromosomes suggests that BIS is a general property of eubacteria. BIS at the various levels may be due to the very similar numbers of codons used in chromosomal halves. Evolutionarily, BIS could have resulted from asymmetric substitution of bases combined with genetic rearrangements. However, the provocative theoretical alternative of whole-genome inverse duplication is here considered.

Keywords: Complementary triplets; Bacterial replication origins; Dot-plot analyses; Codons in leading and lagging strands; Strand asymmetry; Chargaff's DNA parity rules; Bacterial genome evolution

In the classical model for duplication of bacterial chromosomes replication starts at a single site denominated the origin of replication (*ori*) and proceeds bi-directionally until the replication stop site, or replication terminus (*ter*), is met [1]. To a close approximation, *ori* and *ter* are equidistantly located in circular chromosomes. In analogy, in the linear chromosome of *Borrelia burgdorferi* there is a single *ori* located in the middle of the sequence [2,3]. Therefore, in what relates to the replication mechanism, most bacterial chromosomes can be divided into two functional units or replichores [4] and these units are approximately equivalent to chromosomal halves. Because chromosomal halves often differ in base composition the alternative name of chirochores has been suggested for them [5].

During bi-directional duplication, one DNA strand is replicated continuously (leading strand) while the other (lagging strand) is replicated discontinuously [1]. So, the chromosome contains two lagging strands and two leading strands, one of each per chromosomal half (Fig. 1). In general, no distinction is made between these several strands when referring to that, as a rule of thumb, $[A] > [T]$ and $[C] > [G]$ in the lagging strand while $[G] > [C]$ and $[T] > [A]$ in the leading strand [6–9]. However, because of these base disparities the transition between leading and lagging strands in chromosomal halves can be recognised by shifts in $G+C$ contents. Such shifts are often revealed with the formula $[G]-[C]/[G]+[C]$ and they have thus been dubbed GC-skews [6,10].

Differences between leading and lagging strands include a potential greater susceptibility to mutagenesis of the lagging strand [11] and different coding potential because there are more open reading frames (ORFs), and a larger proportion of highly expressed genes, in the leading than in the lagging strand [9]. Codon usage is also different between the leading and lagging strands [12,13].

Inequalities between leading and lagging strands may be due to biased mutation and/or selection associated to transcription [6] or replication [8,9]. Disparities between

[*] Corresponding author. Fax: +52-777-329-7913.
E-mail address: joaquin.sanchez@microbio.gu.se (J. Sánchez).

0006-291X/02/$ - see front matter © 2002 Elsevier Science (USA). All rights reserved.
PII: S0006-291X(02)02583-4

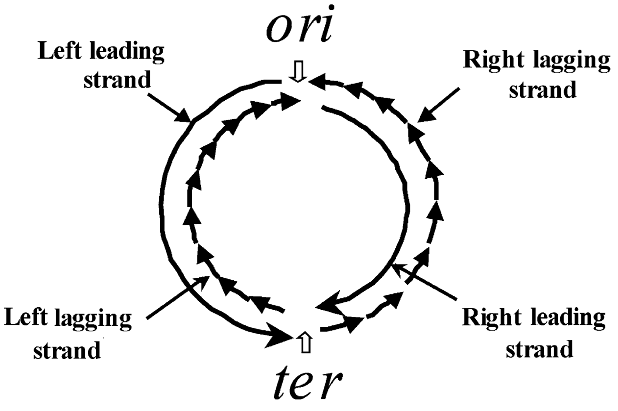

Fig. 1. Diagrammatic representation of leading and lagging strands in chromosomal halves. An idealised circular bacterial chromosome is presented. Relative positions of the chromosomal replication origin (*ori*) and terminus (*ter*) are indicated. Single long arrows are used to indicate continuous replication (leading strand) and multiple short arrows to denote discontinuous replication (lagging strand). To distinguish leading and lagging strands they have been arbitrarily designated as either left or right. Reference in the text to opposing leading and lagging strands should be considered equivalent to: left leading vs. right lagging strands, or left lagging vs. right leading strands, either for circular or linear chromosomes.

the two strands have been considered elements of chromosomal asymmetry [14] and revealing its basis and consequences represents an attractive and challenging subject of research.

Analysis of the distribution of DNA tri-nucleotides (triplets) in the *B. burgdorferi* chromosome revealed parity in complementary triplet frequencies (CTF) between single-stranded chromosomal halves, which are equivalent to opposing leading and lagging strands (Fig. 1). Similar CTF correlations were obtained for 45 other eubacteria. Based on these findings we propose that chromosomes are not compositionally asymmetric but rather that they have bilateral inverse symmetry (BIS). In this paper we primarily explore the basis of BIS and we also discuss its potential evolutionary origins.

Materials and methods

Bacterial chromosomes. Accession numbers (NCBI, GenBank resource from the NIH) for chromosomes whose results are graphically presented are: (alphabetically ordered by species name) NC_000964, NC_001318, NC_002163, NC_000117, NC_000913, NC_000962, NC_002745, NC_002737, and NC_000919.

Software. All sequence manipulations were carried out with an evaluation copy of OMIGA_113(Oxford Molecular Ltd, UK). For convenience, when processing data to generate cumulative position plots (CPP) axes were not used in the classical way but they were inverted so that the dependent variable was in the abscissa and the independent variable in the ordinate. Randomisation of sequences (shuffling) was done with the program Shuffle (http://bcf.arl.arizona.edu/). Dot-plot tests were carried out using an evaluation copy of OMIGA 2.0 (Accelrys). The program Backtranslation (http://www.entelechon.com) was used as an aid for triplet shuffling.

Frequencies of complementary triplets and codon usage determinations. To determine complementary triplet frequency (CTF) sequences in their single-stranded form were divided at the known or proposed *ori* and *ter* [3,10,13–16] so as to produce two segments of equal length (chromosomal halves).

To undertake the same CTF analysis in other bacteria the positions of *ori* and *ter* were needed. However, those have not been established for the large majority of chromosomes. Then we used *dnaA* as a surrogate of *ori* because it has been shown to locate close to this gene [16]. In contrast to *ori*, *ter* does not seem to be associated to any particular gene but as above mentioned, it is usually found opposite to it. Then, chromosomal halves were produced by dividing sequences at *dnaA* and at the position exactly across it. Even though strictly speaking such chromosomal halves should not be considered synonymous with opposing leading and lagging strands until putative *ori* and *ter* are experimentally confirmed, the likelihood that they are the same is very high and we thus use the two terms interchangeably.

For *B. burgdorferi*, CTF were also determined in a sequence (ORF-only sequence) containing all open reading frames (ORFs) but neither intergenic regions nor ribosomal or transfer RNAs. Such ORF-only sequence was created by joining all reported ORFs one after the other as naturally ordered and oriented in the chromosome. For codon usage determinations analogous *in silico* sequences were generated for nine other chromosomes but in this last case ORFs were placed all in their coding-wise orientation (retrieved from NCBI as *.ffn files).

To determine amino acid compositions, compiled protein sequences in pre-computed tables were joined one after the other after all annotations had been deleted.

Results

Distribution of complementary triplets along the B. burgdorferi chromosome and its shuffled controls

The CPP for complementary triplets along one strand of *B. burgdorferi* chromosome generated rhomboidal figures for 26 out of the 32 possible complementary triplet pairs (examples in Fig. 2). These rhombi were generated due to reciprocal relationships in triplet frequencies within each chromosomal half whereby to a low triplet frequency (high slope) corresponded a high frequency (low slope) for its complementary and vice versa. Depending upon how marked the differences in slope were, CPP produced figures that ranged from clear-cut rhombi to semi-parallel lines (Fig. 2).

Because triplet frequencies are influenced by base composition [17] the possibility remained that observed distributions were partially, or even primarily, due to nucleotide compositions along the chromosome. An appropriate procedure to define the importance of base composition is randomisation of sequences via shuffling [18]. The single-stranded *B. burgdorferi* chromosome, with a G + C content of 28.6%, follows Chargaff's [19] second parity rule [A] ≈ [T] and [C] ≈ [G]; however, its chromosomal halves do not: in one half [A] > [T] and [C] > [G] while in the other [A] < [T] and [C] < [G]. To determine the role of these base differences on CPP shuffled controls were prepared (three consecutive

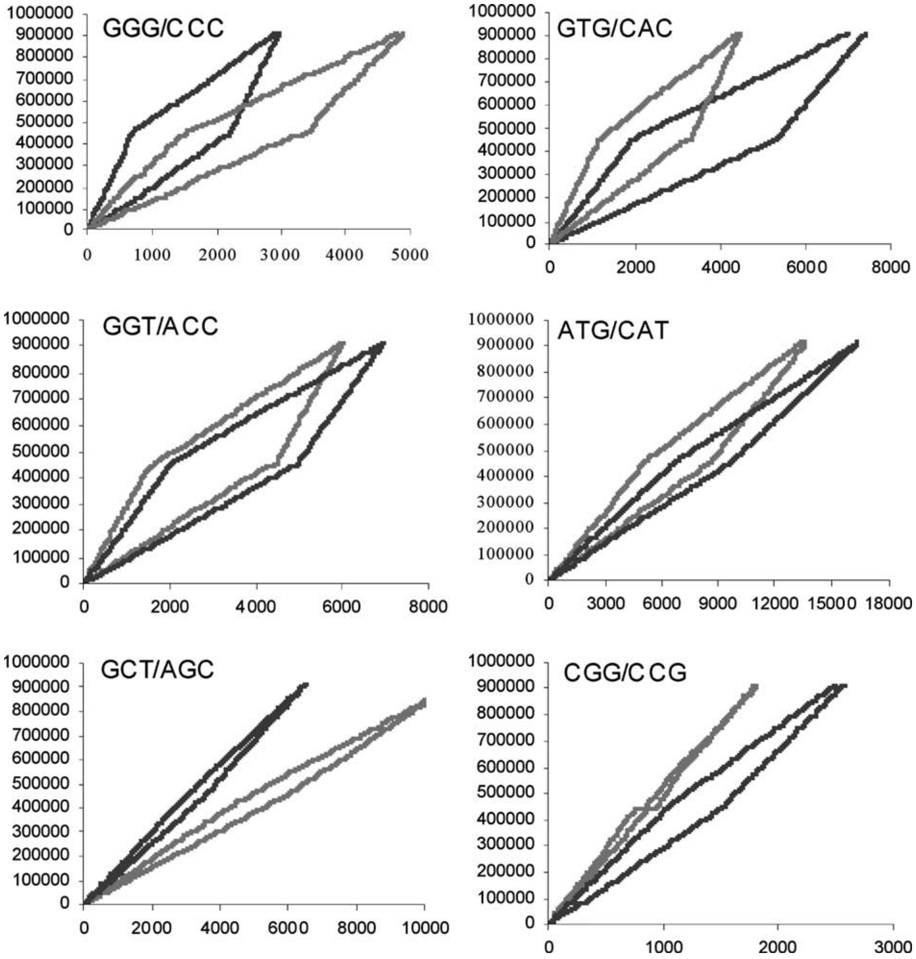

Fig. 2. Representative cumulative position plots for the *B. burgdorferi* chromosome (light line) and the SH shuffled control (dark line). Note that axes are not used in the classical way and the dependent variable (cumulative position value) is in the abscissa while the independent variable (triplet position) is in the ordinate. Axis labels have been removed for clarity. In each panel, analysed triplets are identified (for example, GGG/CCC). These designations also serve to indicate that distributions for the triplet on the left side (e.g., GGG) correspond to the two left sides of rhombi (or to the left semi-parallel line), while distributions for the other triplet (i.e., CCC) correspond to the right sides of rhombi (or right semi-parallel line).

shufflings). The chromosome was either shuffled as a unit (shuffled whole, SW) or each of its halves shuffled separately to then be rejoined (shuffled halves, SH). Neither the SW nor the SH control sequences reproduced the CPP for the original sequence although the SH control generated figures that in general had the same shape as those for the original sequence (Fig. 2). This was not totally unexpected given that base compositions should have determined the global tendencies of both length and slope of the sides of rhombi. Importantly, however, rhombi for the SH control did not have the same length along the x-axis (Fig. 2). Since this length was dependent upon triplet frequency this demonstrated disparities in triplet frequencies with the original chromosome. Meaningfully, CPP could in addition to length differ also in shape, such was the case for CGG/CCG (Fig. 2) and TCG/CGA (not shown) among others. Another significant dissimilarity between the SH control and the original chromosome was that in the former sequence triplets of the same base composition, regardless of their identity, occurred at very similar frequencies (clustering). This produced very similar CPP, even when triplets were unrelated in sequence (Fig. 2, and results not shown). This only very seldom occurred in the original chromosome. Related frequency clusterings have been revealed for other shuffled DNA sequences applying specialised frequency analysis techniques [17].

To find out if differences between the SH sequence and the original chromosome could be explained by variability inherent to the shuffling procedure we carried out repetitive shuffling. Five independently generated SH sequences were obtained and average triplet frequencies and root mean square values (rms) were calculated. The rms values were between 0.1% and 3.4% with a mean of 1%. Therefore, differences between the SH control and the original chromosome may be considered highly significant.

Correlation of complementary triplet frequencies between chromosomal halves

As discussed above, in the CPP for *B. burgdoferi* opposite sides of rhombi represent the distribution of a triplet (or its complementary) in different chromosomal halves and, as for whole rhombi, the length of each side along the *x*-axis was synonymous with triplet frequency. Thus, the observation that opposite sides of rhombi were parallel and of almost identical length indicated parity in CTF between chromosomal halves. This proved to be the case. For instance, ATG and its complementary CAT were found 5064 and 8585 times, respectively, in one-half of the chromosome, while they appeared 8303 (ATG) and 5055 (CAT) times in the other chromosomal half. Analogous parity was found for virtually all triplets in the *B. burgdorferi*. These results demonstrated inverse compositional symmetry at the level of triplets in the *B. burgdorferi* chromosome. Because this symmetry occurs between chromosomal halves we propose that it should be referred to as: bilateral inverse symmetry (BIS).

To investigate if BIS was expressed in other bacterial chromosomes with different G + C contents we examined the chromosomes of *Treponema pallidum* with G + C contents of 52.78%, of four Gram-positives (*Mycobacterium tuberculosis*, *Bacillus subtilis*, *Streptococcus pyogenes*, and *Staphylococcus aureus*) with G + C contents of 65.62%, 43.52%, 38.5%, and 32.8%, respectively, of two Gram-negatives (*Escherichia coli* K-12 and *Campylobacter jejuni*) with G + C contents of 50.79% and 30.54% correspondingly, and of the intracellular bacteria *Chlamydia trachomatis* with G + C contents of 41.3%.

To obtain correlations between complementary triplets the inverse complement of one chromosomal half was first generated to then search for correlations between identical triplets. With this method high correlations for *B. burgdorferi* and for the other nine chromosomes were demonstrated (Fig. 3, filled circles). Therefore, all analysed chromosomes exhibited BIS, irrespective of their very diverse origin and broad differences in G + C contents. To control for the effect of homogeneous CTF within each chromosomal half (internal symmetry) we also correlated triplets without inversely complementing the test chromosomal half (Fig. 3, open circles). As in *B. burgdorferi,* the former usually correlated better than the latter and in the illustrated examples the highest contrasts were seen for *T. pallidum* and *C. trachomatis*.

To further test the universality of BIS we extended CTF analysis to other eubacteria and we found correlations analogous to those in Fig. 3 for: *Bacillus halodurans*, *Buchnera sp.*, *Caulobacter crescentus*, *Clostridium acetobutylicum*, *Chlamydia muridarum*, *Chlamydophila pneumoniae* AR39, *Chlamydophila pneumoniae* CWLO29, *Chlamydophila pneumoniae* J138, *E. coli* 0157:H7, *E. coli* O157:H7_EDL933, *Helicobacter pylori* 26695, *Helicobacter pylori* J99, *Haemophilus influenzae*, *Lactococcus lactis*, *Mycobacterium tuberculosis* CDC1551, *Mycobacterium leprae*, *Mycoplasma genitalium*, *Mycoplasma pneumoniae*, *Mycoplasma pulmonis*, *Neisseria meningitidis* C58, *Neisseria meningitidis* Z2479, *Pasteurella multocida*, *Staphylococcus aureus* MU50,

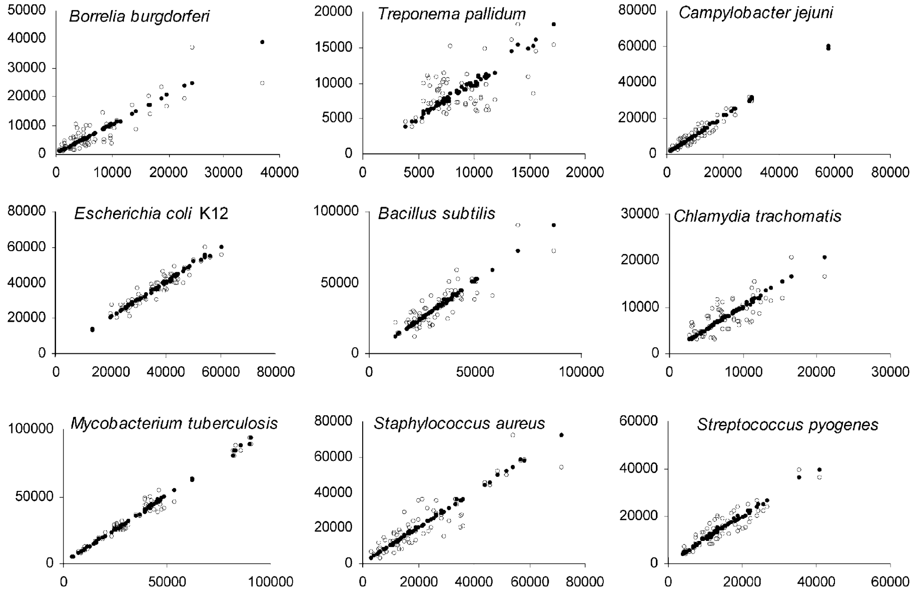

Fig. 3. Correlations in complementary triplet frequencies within and between chromosomal halves of bacteria. For the sake of clarity axis labels have been removed. Triplet frequencies in one half of the single-stranded chromosome (horizontal axis) are plotted against frequencies in the other half (vertical axis) either in its direct form (empty circles) or as a reverse complement (filled circles).

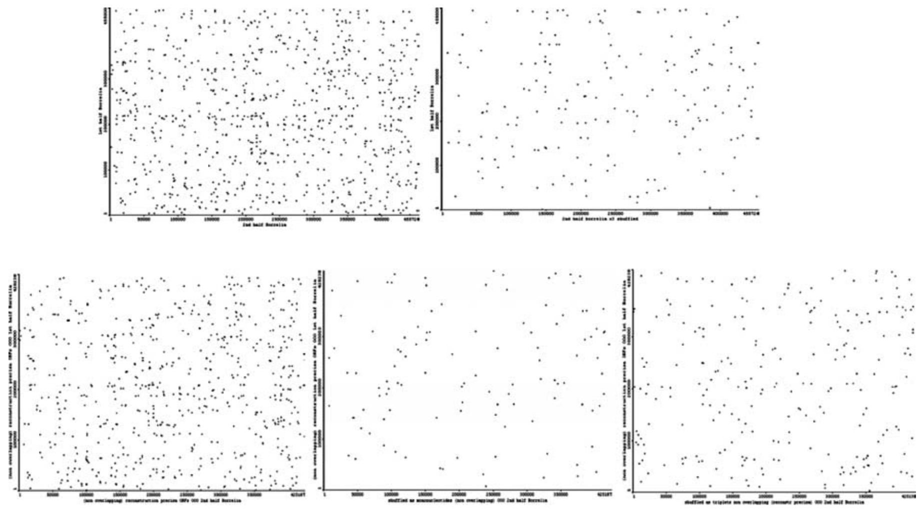

Fig. 4. Demonstration of inverse homology between *B. burgdorferi* chromosomal halves using dot-plot. In the top left panel one chromosomal half was compared to the reverse complement of the other half. In the top right panel: comparisons after shuffling bases in one of the two halves. In the low left panel: comparisons between intact ORFs in chromosomal halves. Low center panel: comparisons after shuffling ORFs in one of the two chromosomal halves at the level of bases. Low right panel: comparisons after shuffling ORFs in one of the two chromosomal halves at the level of triplets (codons).

Streptococcus pneumoniae R6, *Streptococcus pneumoniae* TIGR4, *Ureaplasma urealyticum*, *Vibrio cholerae* chromosomes I and II, and *Yersinia enterocolitica*. To summarise data we calculated the correlation coefficient r^2 for CTF correlations. Including the nine genomes presented in Fig. 3 the average r^2 for CTF correlations between chromosomal halves was 0.993 with a range varying from 0.999 for *C. acetobutylicum* to 0.953 for *V. cholerae* chromosome II.

We also analysed *Aquifex aeolicus*, *Deinococcus radiodurans*, *Mesorhizhobium loti*, *Ricketsia conorii*, *Ricketsia prowazekii*, *Sinorhizobium meliloti*, *Synechocystis* PCC6803, and *Thermotoga maritima*. In these bacteria there were very high CTF correlations between chromosomal halves and thus expression of BIS. However, high CTF correlations *within* chromosomal halves, or high internal symmetry caused that, at difference with the other cases, direct or inverse correlations were essentially indistinguishable.

Dot-plot tests reveal homology between B. burgdorferi chromosomal halves

Because correlations in complementary triplets could indicate inverse sequence similarities between chromosomal halves we searched for inverse homology between *B. burgdorferi* chromosomal halves using a Dot-plot test. With this method, homology was revealed between chromosomal halves with many combinations of software parameters (window, stringency, etc.). Shown results (870 dots, Fig. 4) are for a 35 bp window and 80% stringency. To assess the role of simple probabilistic arrangement of bases over the reported homologies, one of the two chromosomal halves was shuffled. The number of dots obtained using this control sequence was reduced to 210 (Fig. 4). Consequently, results suggested that inverse homology between *B. burgdorferi* chromosomal halves was not just statistical. To determine how much this homology was dependent on protein-coding regions, an ORF-only sequence (Materials and methods) was studied. Equivalents to chromosomal halves from this ORF-only sequence were obtained and dot-plot tests were run for intact sequences or after their shuffling at the level of bases or at the level of triplets (i.e., codons but in their original orientation). Results (Fig. 4) revealed inverse homology also between ORFs in chromosomal halves with the total number of dots amounting to 705. In contrast, shuffling at the level of bases and at the level of codons reduced the number of dots to 135 and 260 dots, respectively (Fig. 4). Lower homologies may not be attributable to variability inherent to the shuffling procedure because standard deviations ($n = 5$) were $\leqslant 5\%$.

Bilateral inverse symmetry and coding contents

To complete the analyses of the role of protein-coding regions we searched for BIS. In the ORF-only sequence correlations between chromosomal halves at the level of complementary bases, dinucleotides, and triplets were virtually identical to those for the intact chromosome (not shown). Thereby, intergenic and RNA sequences contributed little to BIS. This agreed with their low representation (6%) in the chromosome. A more important conclusion from this analysis, however, was that it indicated that codons, and not just triplets, could also follow BIS. This proved to be the case. High correlations between chromosomal halves were obtained

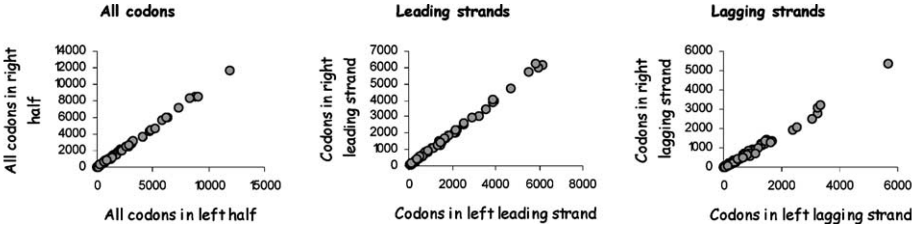

Fig. 5. Correlations in codon contents between *B. burgdorferi* chromosomal halves. Left panel, correlation of all codons. Center panel, correlation of codons in leading strands. Right panel, correlation of codons in lagging strands. Designations of strands as left or right are arbitrary and in accordance with explanations under Fig. 1.

when all codons in each chromosomal half were compared altogether or when codons in leading or lagging strands were analysed separately (Fig. 5).

These results were independently corroborated by determining amino acid compositions of encoded proteins in pre-computed tables at NCBI. Similar analyses of the nine other chromosomes in Fig. 3 demonstrated equally high correlations at the level of both codons and amino acids.

To estimate how important the natural ordering of codons in ORFs was for BIS the ORF-only sequence was shuffled in various ways. We shuffled it either as a unit and at the level of bases (SWB, shuffled whole as bases), or as a unit and at the level of triplets (SWT, shuffled whole as triplets). In addition, halves were shuffled independently and at the level of bases (SHB, shuffled halves as bases), or at the level of triplets (SHT, shuffled halves as triplets). The CPP for the SWB and SWT sequences produced only straight lines in all cases (not shown) while CPP for SHB and SHT generated mostly rhombi. Rhombi for the SHT control were closer (in both shape and length) to those for the ORF-only sequence (GGG/CCC is given as example in Fig. 6).

Hence, the arrangement of codons in ORFs was of importance for BIS at the level of triplets but codon contents alone appeared to already be a major determinant of it. In contrast, base compositions had a lesser influence over BIS and this reflected in the more sizeable differences between CPPs for SHB and the unshuffled sequence (Fig. 6).

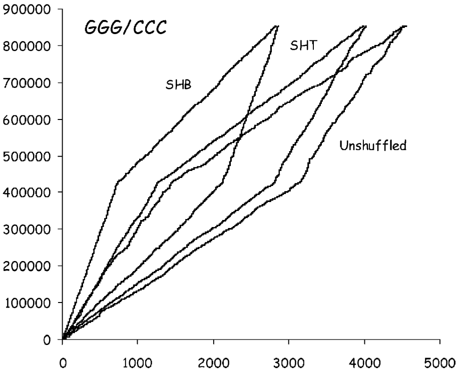

Fig. 6. Cumulative position plots for GGG and CCC in a *B. burgdorferi* ORF-only sequence either in its intact form (unshuffled) or after shuffling of ORFs in chromosomal halves at the level of bases (SHB), or at the level of triplets (SHT). Axes are as for Fig. 2.

Discussion

Bilateral inverse symmetry, or BIS, in bacterial chromosomes has not been previously suggested; rather, chromosomes are regarded as asymmetric on the basis of base composition differences between leading and lagging strands. Nonetheless, an observation has been made that base skews in *E. coli* chromosomal halves almost exactly cancel out each other [20] and such observation is consistent with the concept of BIS. Actually, if leading and lagging strands in each chromosomal half (Fig. 1) are separately considered BIS at the level of individual bases and up to at least the level of triplets is quickly revealed.

BIS at the single-nucleotide level would be equivalent to the so-called Chargaff's second parity rule ([A] $\approx$ [T] and [C] $\approx$ [G] in single-stranded DNA) [19,21] but in a form that is applicable only to whole bacterial chromosomes. Practically all so-far sequenced bacterial chromosomes follow this rule. However, we found that in at least 39 bacteria-representing 36 different species-chromosomal halves did not. In chromosomal halves compositions were so that in general [A] $\neq$ [T] and [C] $\neq$ [G] with deviations not always in the classical direction for leading and lagging strands. Importantly, independently of tendencies, in virtually every case base composition differences practically exactly cancelled each other out. Thereby, BIS at the single-nucleotide level would explain Chargaff's second parity rule as applicable to whole bacterial chromosomes.

Since in principle higher-order level BIS could be just a consequence of single-base BIS we explored this possibility in the *B. burgdorferi* chromosome. Using shuffled control sequences we demonstrate that BIS at the single-nucleotide level does not determine higher-order BIS, at least not as expressed by the original chromosome. This may be the case for many, if not all, eubacteria because analogous results have been obtained for *C. acetobutylicum*, *C. trachomatis*, *S. pyogenes*, *Buchnera sp.*, and *E. coli* (unpublished).

To explore the maximal *n*-tuple size at which BIS would still express in the intact *B. burgdorferi* chromosome, we analysed 156 out of the 256 possible 4-tuples and found that almost all distributed according to BIS. However, analysis of several randomly chosen 5-tuples strongly suggests that BIS disrupts at that point. Nevertheless, when *n*-tuples were imbalanced in base composition they distributed differently. For example, C-only and T-only sequences up to pentaplets and octamers, respectively, still follow BIS in *B. burgdorferi*. This applied also to many T-rich octamers including the *ori*-biased [17] TTGTTTTT sequence. We found an analogous situation in *E. coli* for the G-rich GCTGGTGG Chi octamer, the *ori*-biased signal with a role in recombination associated to chromosomal replication [22,23].

Symmetrical DNA features related to those here presented have previously been found [21,24,25] but their biological origins, or consequences, have not been determined. Our results hint on a relation with the coding potential of DNA because a *B. burgdorferi* sequence containing only ORFs expressed BIS even after codons were shuffled.

A correlation in codon contents between chromosomal halves implies that basically identical sets of nucleotides, amino acids, tRNAs, etc., will be used when transcribing and translating each chromosomal half. Remarkably, this striking genetic organisation occurs even though halves encode for different proteins. Such genetic settings were not a special case for *B. burgdorferi* and codon content correlations between chromosomal halves were found in at least nine other bacteria. Consequently, this genomic condition may be a *sine qua non* for bacterial chromosomes and as such perhaps be the reason for the widespread occurrence of BIS.

Connected to the above findings, and in order to address the potential evolutionary origin of BIS, two proposals regarding DNA properties of bacterial species may be considered: (a) nucleotide compositions of bacterial chromosomes seem to be fixed within borders that are characteristic of the species, possibly as a result of competition for metabolic resources [26] and (b) frequencies of di-nucleotides in bacteria may also be set within limits and in a species-specific manner [27]. Then, upon construction of the chromosome, and/or maintenance of its stable structure, neither base compositions nor base arrangements would seem to occur in a fully free manner. These limitations would most likely have a direct impact on codon frequencies along the chromosome. Thus, one could propose that one consequence of these restrictions may be BIS itself.

Compositional limitations could have been produced during evolution through the asymmetric substitution of bases [5,6] in the leading and lagging strands. Accordingly, substitution of bases could operate so that changes in chromosomal halves turned essentially equal (statistically). Therefore, when a base was changed in the left lagging strand the same change would take place somewhere along the right lagging strand. Alternatively, if opposing leading and lagging strands were implicated, changes would involve not the same but complementary bases. Either form of substitution could have created BIS at the single-nucleotide level.

If asymmetric substitution of bases was coordinated not for individual bases but for small sets of nucleotides, in principle at least, replacements could have created BIS also at the level of triplets and maybe even at higher-order levels.

It should be noted, however, that recent theoretical analyses suggest that asymmetric substitution of bases alone may not generate higher-order symmetry of the DNA double strand [21]. Adapted to the present analysis this suggestion would imply that asymmetric substitution of bases might not be sufficient to account for higher-order BIS. Therefore, asymmetric substitution of bases would have to be combined with genetic rearrangements to generate BIS. This may be possible because the two phenomena should be fully compatible.

However, in *B. burgdorferi*, we found that BIS expressed also in the form of inverse homology that included ORFs. Inverse homology could have been fortuitously produced during asymmetric substitution of bases. However, simple fortuitous combinations would seem to disagree with the drastic reductions in homology caused by shuffling both at the level of bases and at the level of codons. An alternative explanation for the found homology is that it signals the presence of preexistent sequence similarity between opposing leading and lagging strands.

If the latter was so, an alternative evolutionary hypothesis would naturally emerge, which could either be independent of the one above or even be its complement.

It may be possible that BIS, and especially higher-order level BIS, is a vestige of kinship between chromosomal halves. Hence, one could propose that one half of the chromosome begat the other via an ancient whole-genome inverse duplication event.

Such model would not be without precedent as whole-genome duplications have been suggested as a likely pathway by which the *E. coli* chromosome could have been generated [28].

There would be elements both for and against such proposal.

From the theoretical point of view, several chromosomal features could be highly compatible with such model and in particular with one involving whole-genome inverse duplication from *ori* to *ter* (or vice versa). Such type of duplication could first of all explain the occurrence and position of GC-skews. Accordingly, if the original genome violated Chargaff's second parity rule (as occurs, for example, for the very large majority of bacterial plasmids and phages) its inversion and duplication would have immediately produced a base composition skew at the junction between genomes. Such skew would have been centrally located in linear chromosomes. For circular chromosomes two exactly opposing skews would have been created: one at *ori* and another at *ter*. This is precisely how skews occur in current linear and circular bacterial chromosomes.

Upon whole-genome duplication *ori* itself would have been copied but in its inversely complemented form and this would have upturned its functional orientation. This would have created two opposing *ori* and thus two divergently moving replication forks. This could be the origin of bi-directional replication as found in present-day bacterial chromosomes.

Against the model two theoretical evolutionary arguments could be raised: a whole-genome inverse duplication seems conceivable only at the early stages of evolution, and then it appears improbable that traces of such duplication (i.e., BIS) would still remain. On the other hand, the widespread occurrence of BIS would imply that many bacteria underwent whole-genome duplication in an independent manner and relatively recently. Under present evolutionary theories this condition seems unlikely. At the DNA sequence level the distribution of paralogous genes did not seem to agree with a whole-genome inverse duplication hypothesis [29]. Whether the distribution of paralogous genes is connected to BIS will only be solved through an in-depth analysis of the reasons for inverse homology between ORFs in chromosomal halves.

Objections to the model could appear to be strengthened by the fact that no whole-genome inverse duplication has so far been documented in bacteria. Nevertheless, very large stable inverse duplications in *E. coli* have recently been obtained in the laboratory [30]. In fact, such genetic events may be neither unnatural nor uncommon given that it is precisely of large inverse duplications that the so-called amphimeric genomes are made of [31], and such genomes occur in a wide variety of organisms, ranging from archaebacteria to mammals [31]. Furthermore, it is exactly through whole-genome inverse duplication that some bacterial genomes might naturally replicate [32].

Acknowledgments

Both authors acknowledge Sistema Nacional de Investigadores (SNI), CONACyT, México, for stipends. Author J.S. thanks SILANES, S.A. de C.V., México for a financial donation. Author M.V.J. received Grant PAPIIT-IN213199, UNAM, México. Norma Martínez and Juan R. Bobadilla are gratefully acknowledged for help with data processing and Imelda López Villaseñor and Enrique Merino for constructive criticism of the manuscript. We thank Joaquín Sánchez Lugo for valuable language suggestions. We especially thank Eugene Koonin for his kind insightful comments and suggestions on the manuscript.

References

[1] B. Lewin, Genes, vol. VII, Oxford University Press, New York, 2000.

[2] C.M. Fraser et al., Genomic sequence of a Lyme disease spirochaete, *Borrelia burgdorferi*, Nature 390 (1997) 553–555.

[3] M. Picardeau, J.R. Lobry, B.J. Hinnebusch, Physical mapping of an origin of bi-directional replication at the center of the *Borrelia burgdorferi* linear chromosome, Mol. Microbiol. 32 (1999) 437–445.

[4] F.R. Blattner et al., The complete genome sequence of *Escherichia coli* K-12, Science 277 (1997) 1453–1474.

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

[6] M.P. Francino, H. Ochman, Strand asymmetries in DNA evolution, Trends Genet. 6 (1997) 240–245.

[7] M. McLean, K.H. Wolfe, K.M. Devine, Base composition skews, replication orientation, and gene orientation in 12 prokaryote genomes, J. Mol. Evol. 47 (1998) 691–696.

[8] A.C. Frank, J.R. Lobry, Asymmetric substitution patterns: a review of possible underlying mutational or selective mechanisms, Gene 238 (1999) 65–77.

[9] E.R. Tillier, R.A. Collins, The contributions of replication orientation, gene direction, and signal sequences to base-composition asymmetries in bacterial genomes, J. Mol. Evol. 50 (2000) 249–257.

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

[11] X. Veaute, R.P. Fuchs, Greater susceptibility to mutations in lagging strand of DNA replication in *Escherichia coli* than in leading strand, Science 261 (1993) 598–600.

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

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

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

[15] E.P. Rocha, A. Danchin, A. Viari, Universal replication biases in bacteria, Mol. Microbiol. 32 (1999) 11–16.

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

[17] D.R. Forsdyke, Relative roles of primary sequence and (G + C)% in determining the hierarchy of frequencies of complementary trinucleotide pairs in DNAs of different species, J. Mol. Evol. 41 (1995) 573–581.

[18] M. Gribskov, J. Deveraux, Sequence Analysis Primer, New York Stockton Press, New York, 1991.

[19] E. Chargaff, How genetics got a chemical education, Ann. N. Y. Acad. Sci. 325 (1979) 345–350.

[20] C. Shioiri, N. Takahata, Skew of mononucleotide frequencies, relative abundance of dinucleotides, and DNA strand asymmetry, J. Mol. Evol. 53 (2001) 364–376.

[21] P.F. Baisnee, S. Hampson, P. Baldi, Why are complementary DNA strands symmetric?, Bioinformatics 18 (2002) 1021–1033.

[22] D.A. Dixon, S.C. Kowalcykowski, The recombination hotspot Chi is a regulatory sequence that acts by attenuating the nuclease activity of the *E. coli* RecBCD enzyme, Cell 73 (1993) 87–96.

[23] R. Uno, Y. Nakayama, K. Arakawa, M. Tomita, The orientation bias of Chi sequences is a general tendency of G-rich oligomers, Gene 259 (2000) 207–215.

[24] Q. Dong, A.J. Cuticchia, Compositional symmetries in complete genomes, Bioinformatics 17 (2001) 557–559.

[25] V.V. Prabhu, Symmetry observations in long nucleotide sequences, Nucleic Acids Res. 12 (1993) 2797–2800.

[26] E.P. Rocha, A. Danchin, Base composition bias might result from competition for metabolic resources, Trends Genet. 18 (2002) 291–294.

[27] H. Nakashima, M. Ota, K. Nishikawa, T. Ooi, Genes from nine genomes are separated into their organisms in the dinucleotide composition space, DNA Res. 30 (1998) 251–259.

[28] D. Zipkas, M. Riley, Proposal concerning mechanism of evolution of the genome of *Escherichia coli*, Proc. Natl. Acad. Sci. USA 72 (1975) 1354–1358.

[29] J.A. Eisen, J.F. Heidelberg, O. White, S.L. Salzberg, Evidence for symmetric chromosomal inversions around the replication origin in bacteria, Genome Biol. 1 (2000) 1–9.

[30] N. Handa, Y. Nakayama, M. Sadykov, I. Kobayashi, Experimental genome evolution: large-scale genome rearrangements associated with resistance to replacement of a chromosomal restriction-modification gene complex, Mol. Microbiol. 40 (2001) 932–940.

[31] E. Rayko, Organization, generation and replication of amphimeric genomes: a review, Gene 199 (1997) 1–18.

[32] K. Kobryn, G. Chaconas, The circle is broken: telomere resolution in linear replicons, Curr. Opin. Micro. 4 (2001) 558–564.