J Mol Evol (2001) 53:615–621
DOI: 10.1007/s002390010248

JOURNAL OF MOLECULAR EVOLUTION

The Differential Killing of Genes by Inversions in Prokaryotic Genomes

Paweł Mackiewicz,[1] Dorota Mackiewicz,[1] Agnieszka Gierlik,[1] Maria Kowalczuk,[1] Aleksandra Nowicka,[1] Małgorzata Dudkiewicz,[1] Mirosław R. Dudek,[2] Stanisław Cebrat[1]

[1] Institute of Microbiology, Wrocław University, ul. Przybyszewskiego 63/77, 51-148 Wrocław, Poland
[2] Institute of Physics, Zielona Góra Pedagogical University

Received: 12 December 2000 / Accepted: 20 April 2001

Abstract. We have elaborated a method which has allowed us to estimate the direction of translocation of orthologs which have changed, during the phylogeny, their positions on chromosome in respect to the leading or lagging role of DNA strands. We have shown that the relative number of translocations which have switched positions of genes from the leading to the lagging DNA strand is lower than the number of translocations which have transferred genes from the lagging strand to the leading strand of prokaryotic genomes. This paradox could be explained by assuming that the stronger mutation pressure and selection after inversion preferentially eliminate genes transferred from the leading to the lagging DNA strand.

Key words: Gene inversion — Mutation pressure — Ortholog — DNA asymmetry — Leading — Lagging strand

Introduction

There are many asymmetric processes which treat DNA strands unequally. The most effective of them are mutational pressures associated with replication and transcription, and uneven distribution of coding and signal sequences. For review see: Francino and Ochman 1997; Mrazek and Karlin 1998; Frank and Lobry 1999; Karlin 1999. As a result of asymmetrical processes acting on DNA, bias in the base and codon composition of DNA strands has been observed in many bacterial genomes (e.g. Lobry 1996; Freeman et al. 1998; Grigoriev 1998; McInerney 1998; McLean et al. 1998; Mrazek and Karlin 1998; Salzberg et al. 1998; Lafay et al. 1999; Mackiewicz et al. 1999; Rocha et al. 1999).

Replication-associated mutational pressure seems to exert the strongest effect introducing bias in the base-composition between the two complementary DNA strands of bacterial genomes (Mackiewicz et al. 1999; Tillier and Collins 2000a). For the first time it has been suggested by Filipski (1990) for virus genomes, and by Lobry (1996) for bacterial genomes. The DNA molecule is built of two antiparallel strands. One of the strands is synthesized continuously (it is called the leading strand) and the other one, called the lagging strand, is synthesized by joining Okazaki fragments (Kornberg and Baker 1992). The differences in replication of the two strands, enzymological and architectural asymmetry of the replication fork, different processivities of polymerases, different error rates, and effectiveness of repair systems implicate different mutational pressures on the two DNA strands. Higher rates of the introduction of errors into the lagging strand has been shown experimentally (e.g. Trinh and Sinden 1991; Basic-Zaninovic et al. 1992; Veaute and Fuchs 1993; Roberts et al. 1994; Iwaki et al. 1996; Thomas et al. 1996). However, Fijałkowska et al. (1998) found that the mutation rate in gene *lacZ* of the *Escherichia coli* lactose operon is higher when inserted into the leading strand than when incorporated into the lagging strand.

In genetic terminology the location of the sense strand

Correspondence to: S. Cebrat; *email:* cebrat@microb.uni.wroc.pl; *website:* http://smORFland.microb.uni.wroc.pl

determines the described location of a gene. If the sense strand is located on the leading strand, it is assumed that "the gene lies on the leading strand." In prokaryotic genomes, the leading and lagging roles of DNA strands are predetermined by location of the origin of replication and the terminus of replication. Thus, a gene stays at the same position in respect to the leading or lagging strand as long as it is not inverted within the same replichore or until it is translocated without inversion into the other replichore (replichores are the two halves of the genome which are replicated in opposite directions, Blattner et al. 1997). In this paper we use the term *inversion* to mean when the position of a gene is switched from the leading to the lagging strand or in the opposite direction. Since it has been found that rearrangements and gene shuffling in bacterial genomes are very frequent (Mushegian and Koonin 1996; Tatusov et al. 1996; Kolsto 1997; Watanabe et al. 1997; Bellgard et al. 1999; Itoh et al. 1999), it seems important to analyze gene translocations regarding the asymmetrical organization of bacterial chromosomes and the leading and lagging roles of DNA strands.

There is still no consensus about which strand—leading or lagging—is synthesized with higher fidelity. We have found that the divergence rate of genes located on the lagging strand is statistically significantly higher than that of genes located on the leading strand (Szczepanik et al. 2001). However, a lower divergence rate is not a direct indication that the mutation rate on the leading strand is lower—it is possible that selection for genes lying on the leading strand is stronger. But if we assume that selection is responsible for the more conserved character of the genes on the leading strand, we should ask: why are these genes located preferentially on the leading strand? It seems reasonable to answer that it is because the mutation pressure on the genes located on this strand is lower. It has been recently shown that an inversion of a gene in respect to the leading/lagging strand causes very fast accumulation of mutations in the very first period after inversion (Tillier and Collins 2000b; Szczepanik et al. 2001). McInerney (1998) concluded from his results that there should exist selective advantage of transpositions of highly expressed genes to the leading strand. This statement takes into account the observation that replication proceeds more slowly through a gene that is transcribed in the opposite direction to the replication fork movement (its sense is located on the lagging strand) (French 1992) because of head-on collisions between DNA and RNA polymerases (Brewer 1988). Therefore, transcription of lowly expressed genes would not interfere with replication and interruption of their transcription should not be as deleterious as of highly expressed genes. If it is true that the more conserved genes located on the leading strand are more sensitive to mutations and, when mutated, they could be more frequently killed (eliminated) by selection than the genes located on lagging strand, the specific bias in the results of inversion processes should be observed. If we assume that the number of inversions of genes located on the leading and lagging strands should be proportional to the number of these genes located at a specific DNA strand, the difference between the expected number and the number of "accepted" translocations should reflect the bias in the killing of genes by selection. Thus, we should observe relatively more genes which were transferred from the lagging to the leading strand than genes which have been moved in the opposite direction. To show this, we have elaborated a method which enables us to determine the direction in which the gene has changed its position in respect to the role of the DNA strand during replication.

Materials and Methods

Data for Analysis

All analyses have been done on two pairs of genomes: *Chlamydia trachomatis* (Stephens et al. 1998) versus *C. pneumoniae* (Kalman et al. 1999), and *Borrelia burgdorferi* (Fraser et al. 1997) versus *Treponema pallidum* (Fraser et al. 1998). Their sequences have been downloaded from ftp://www.ncbi.nlm.nih.gov. The two pairs of genomes were chosen because, according to parameters of asymmetry in each of these genomes, the genes lying on the leading strand form a set distinct from the set of genes lying on the lagging strand.

We have extracted amino acid sequences of orthologs of the studied genomes from Clusters of Orthologous Groups (COGs) downloaded 20 January 2000 from ftp://www.ncbi.nlm.nih.gov/pub/COG. COGs contain proteins which are supposed to have evolved from one ancestral protein (Koonin et al. 1998; Tatusov et al. 2000). In the construction of COGs the authors have used the best-hit rule, but not an arbitrarily chosen statistical cut-off value. This approach accommodates both slow- and fast-evolving proteins and makes COGs useful for evolution analyses. Orthologs extracted from COGs were identified with ORFs annotated in data bases of analyzed genomes.

For each pair of organisms, orthologs were classified into three groups according to their location on the DNA strand: sequences lying on leading strands in both compared genomes, sequences lying on lagging strands in both compared genomes, and sequences which changed the strand—lying on the leading strand in one genome and on the lagging strand in the other genome. Boundaries between leading and lagging strands (position of origin and terminus of replication) and decisions concerning the location of genes on one of these strands were set on the basis of the results of DNA walks describing nucleotide compositional bias of DNA strands (Mackiewicz et al. 1999, see also: http://smorfland.microb.uni.wroc.pl).

Estimation of the Rate of Divergence (Evolutionary Distances) of Orthologs

Amino acid sequences of COGs of the analysed genomes were aligned by the CLUSTAL W 1.8 v. program (Thompson et al. 1994). To estimate evolutionary distances, pairwise distances (expressed by the mean number of amino acid substitutions per site) between sequences of each COG were calculated with the program PROTDIST, from the PHYLIP 3.5c package (Felsenstein 1993) using a model based on the Dayhoff PAM substitution matrix (Dayhoff et al. 1978). Only the closest orthologs in each COG have been analyzed.

For each group of orthologs (sequences lying on leading strands in

both genomes, lying on lagging strands, and the ones which changed their positions), the mean value of the evolutionary distances was counted. The statistical significance of differences between these values was estimated with ANOVA Kruskal–Wallis test (Sokal and Rohlf 1995).

Determination of the Direction of Gene Transfer Between the Leading and Lagging Strands

For all four genomes the distributions of all ORFs annotated in data bases in the two dimensional space were prepared (Figs. 1 and 2). In these distributions each gene is represented by a point whose X axis coordinate represents the value of the AT skew [(A − T)/(A + T)] and Y axis coordinate represents the value of the GC skew [(G − C)/(G + C)] of the third codon positions of the gene. The AT skew and GC skew values proved to be good parameters describing asymmetry of DNA strands (Lobry 1996).

For each of the analysed genomes, two distinct sets representing the genes from leading and lagging strands have been obtained. For each of these sets the coordinates of their centres of distributions have been calculated. The coordinates of the centers are the mean values of AT skew and GC skew of all genes in a given set. For each set the standard deviation of the distances of genes from the distribution center have been calculated.

We have taken all pairs of orthologs included in COGs, represented in both genomes of the analyzed pair which are localized on different (leading/lagging) DNA strands. For each ortholog its distances (Euclidean distances) from the two centers of distributions (D) of the two sets (leading and lagging strand genes) have been calculated separately as follows:

$$D = \sqrt{D_x^2 + D_y^2} \quad (1)$$

where: D—distance of the given ortholog to the centre of distribution of the given set,

$$D_x = (x - X)/SD_X, \quad (2)$$

$$D_y = (y - Y)/SD_Y, \quad (3)$$

X and SD_X—mean value and standard deviation of AT skew of the given set,

Y and SD_Y—mean value and standard deviation of GC skew of the given set,

x and y—AT and GC skew values of the analyzed ortholog.

To decide which gene of a given ortholog pair switched its strand, we have compared their distances of the centers of sets of genes lying on leading and lagging strands. We have assumed that if a gene stays for a long time in the same position in respect to the leading/lagging strand, its nucleotide composition fits better to the set it belongs to and it is closer to center of distribution of this set. If a gene has switched its position recently, its distance to the center of its "new" set is greater. Basing on this assumption we have checked the condition:

$$D^1_{own}/D^1_{other} > D^2_{own}/D^2_{other} \quad (4)$$

where: D^1_{own} and D^1_{other}—distance of the analyzed gene to the distribution centre of the gene set of its own and the other strand, respectively.

D^2_{own} and D^2_{other}—distance of the counterpart ortholog of the analyzed gene to the distribution centre of the gene set of its own and the other strand respectively in the other genome.

If the above condition is true, the analysed gene has switched its position, if not—its ortholog has.

For example (see Fig. 2), gene BB0203 from the *B. burgdorferi* genome, lying on the lagging strand of the genome, is very close to the gene set lying on this very strand. Its ortholog—gene TP0114 lying on the leading strand of the *T. pallidum* genome is closer to the centre of gene distribution from lagging strand than to the center of gene distribution of its own strand. According to parameters of asymmetry, gene TP0114 D_{own}/D_{other} parameter is larger than that for its ortholog. Thus, it is most probable that gene TP0114 was transferred from the lagging to the leading strand recently and it still "remembers" compositional properties of the previous strand.

Table 1. The mean evolutionary distances (±SE) between orthologs classified into three sets according to their location on the DNA strand

Sets of orthologs	*B. burgdorferi* vs. *T. pallidum*	*C. pneumoniae* vs. *C. trachomatis*
Lying on leading strand	1.28 (±0.04)	0.43 (±0.02)
Lying on lagging strand	1.39 (±0.08)	0.49 (±0.02)
Which switched strand	1.48 (±0.07)	0.59 (±0.05)

The values are the mean numbers of amino acid substitutions per site between two genomes for each set of orthologs (counted on the base of the Dayhoff PAM matrix model).

Results and Discussion

Since the leading and lagging strands are exposed to different mutational pressures, inversion of a gene from one strand to the other should lead to a very strong mutational pressure in the very first period after inversion. In Table 1 we have compared rates of divergence in three sets of orthologs:

- both orthologs lying on the leading strand,
- both orthologs lying on the lagging strand, and
- the set encompassing pairs of orthologs of which one has switched its position with respect to leading/lagging DNA strand since the two genomes diverged.

The three sets of analyzed orthologs for each pair of genomes statistically significantly differ in the rate of evolution when analyzed by the ANOVA Kruskal–Wallis test ($p < 0.02$). Data clearly show that the divergence in the group which switched their positions is higher than among the other two groups. We have found that it is true also for other pairs of closely related genomes (Szczepanik et al. 2001). This explains the results of Fijałkowska et al. (1998), who found that gene *lacZ* of the *E. coli* lactose operon cumulates more mutations when incorporated into the leading strand than when incorporated into the lagging strand. The usual location of the lactose operon in the *E. coli* genome is the lagging strand. Thus, the effect of inversion is observed, and the general conclusion of the authors that there is a universal rule that the replication-associated mutation rate is higher for genes lying on the leading strand seems wrong. We have found that the divergence rate for genes lying on the leading strand is lower than that for genes lying on the lagging strand for almost all analyzed genomes.

For our present studies we have chosen the pairs: *C. trachomatis* versus *C. pneumoniae,* and *B. burgdorferi*

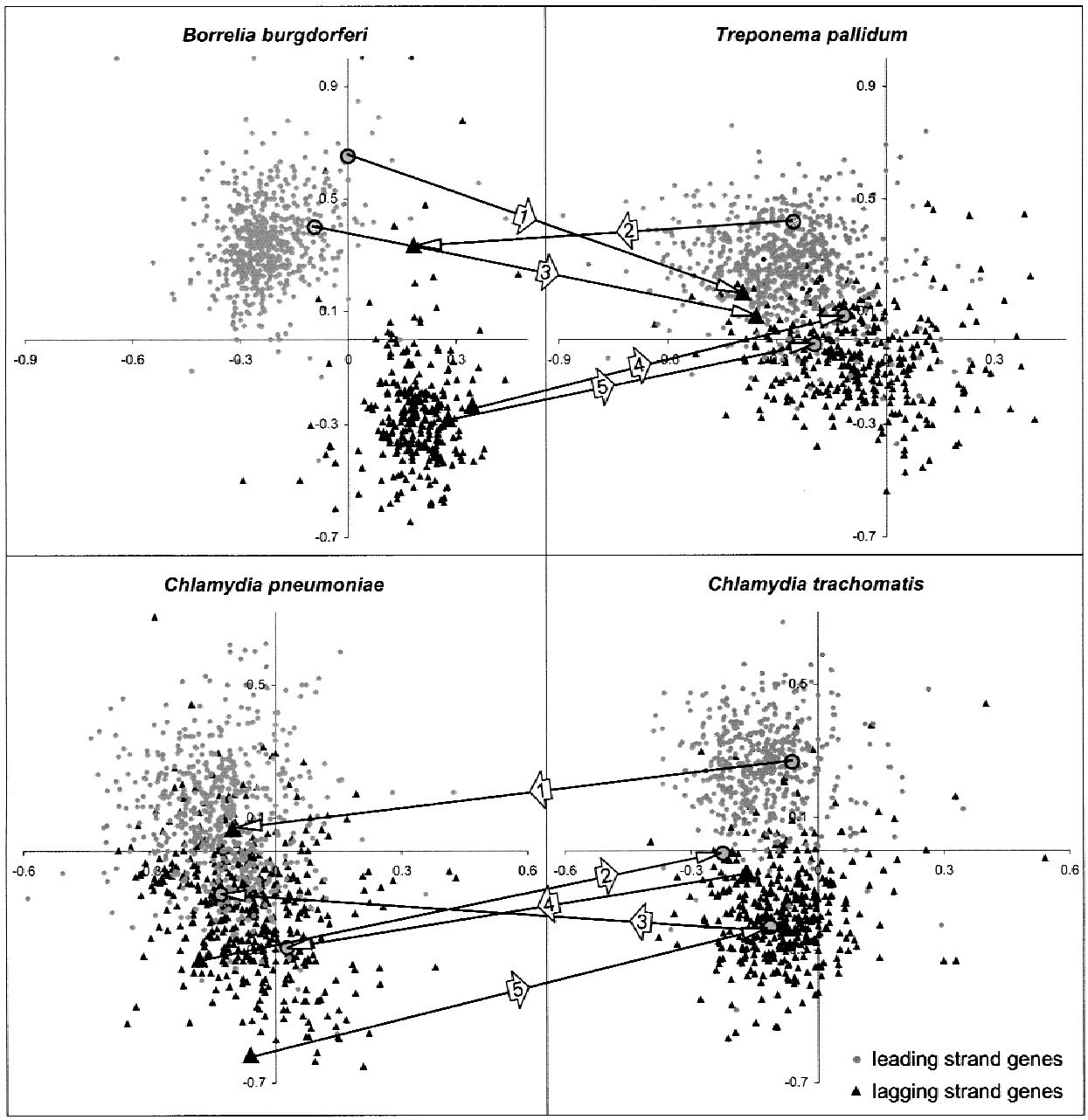

Fig. 1. The distributions of genes from leading (circles) and lagging (triangles) strands in the genomes: *B. burgdorferi*, *T. pallidum*, *C. trachomatis*, and *C. pneumoniae*. Coordinates represent the AT and GC skew values in the third codon positions of the analyzed genes. Arrows connect examples of pairs of orthologs which are localized on different (leading/lagging) DNA strands. Arrow-head points to the ortholog which most likely has switched the strand (according to the procedure described in the Methods section). Pairs of orthologs between *B. burgdorferi* and *T. pallidum* used as examples are: (1) BB0696–TP0906, (2) BB0551–TP0366, (3) BB0025–TP0474, (4) BB0450–TP0111, (5) BB0263–TP0926. Pairs of orthologs between *C. pneumoniae* and *C. trachomatis* used as examples are: (1) CPn0720–CT659, (2) CPn0263–CT221, (3) CPn0867–CT726, (4) CPn0956–CT805, (5) CPn0140–CT212.

versus *T. pallidum* because we were able to divide the sets of genes of these organisms into two distinct subsets—those lying on leading strands and those lying on lagging strands.

In Fig. 1 the distributions of genes in the space with coordinates representing the AT and GC skew values in the third codon positions are presented. In each of the analysed genomes genes from leading and lagging strands form two distinct sets. Only for *Chlamydia* genomes these sets partially overlap.

It is possible to recognize the orthologs which have changed strand (according to the procedure described in Methods section and Fig. 2). In Fig. 1 arrows connect examples of pairs of orthologs which are localized on different (leading/lagging) DNA strands. Arrow-head points to the ortholog which most likely has switched the

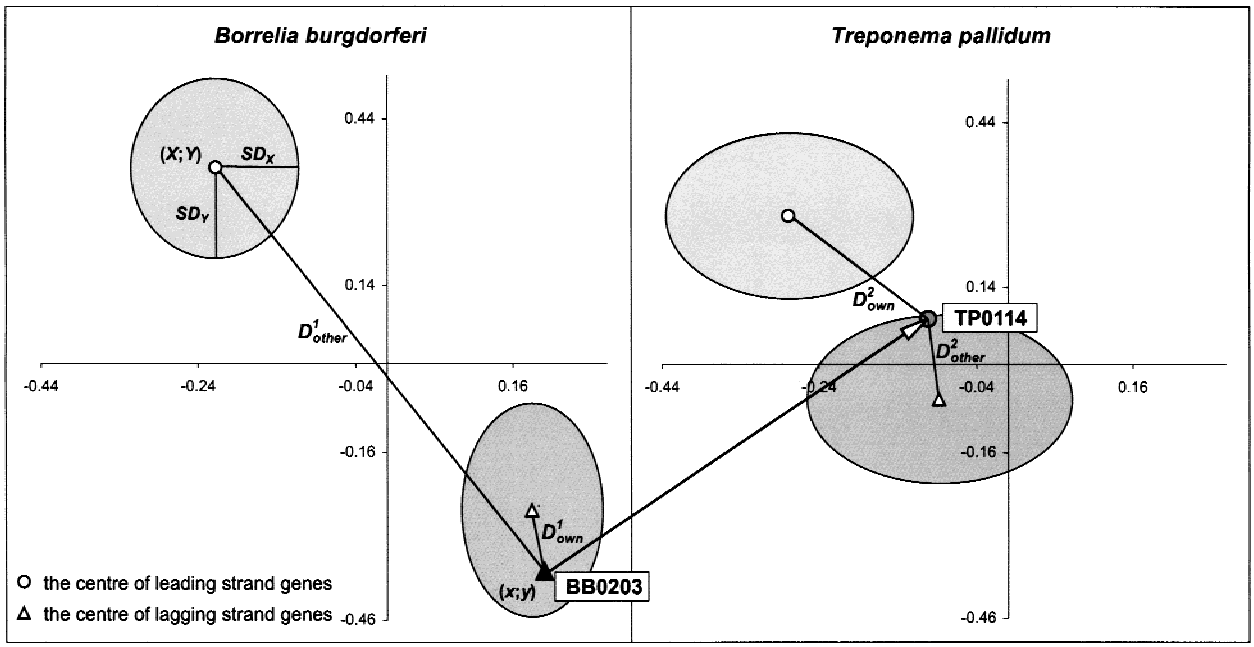

Fig. 2. Graphical presentation of the method of determining the direction of the orthologs transfer between leading and lagging strands. X axis shows the values of the AT skew [(A − T)/(A + T)] and Y axis—the values of the GC skew [(G − C)/(G + C)] of the third codon positions of analyzed genes. Ellipses represent sets of genes lying on leading and lagging strands in the *B. burgdorferi* and *T. pallidum* genomes. The centers of ellipses (*X, Y*) correspond to mean values of AT and GC skew of the given gene set. Standard deviations of AT and GC skew (SD_X, SD_Y, respectively) correspond to the length of the ellipses' semi-axes. One pair of orthologs found in the *B. burgdorferi* (gene BB0203) and *T. pallidum* (gene TP0114) genomes has been shown (connected by an arrow). For each ortholog its distances (Euclidean) from the center of the gene set of its own (D_{down}) and the other strand (D_{other}) have been calculated as described in the text.

strand. There are some pairs of orthologs of which one gene switched its position very recently (i.e. TP0114 in *T. pallidum* genome in Fig. 2) and "remembers" the nucleotide composition of the third codon positions characteristic for its previous position. Some orthologs possess intermediate values of AT and GC skew and are located between centers of distributions of the two sets. They switched their strands earlier and have had more time to accumulate mutations so they become more similar to the genes of the new strand. Then, the distance to the center of genes set of the new strand seems to be negatively correlated with the time which they have spent on it already. In fact, we have found such negative correlation (data not shown).

The distributions of genes between the leading and lagging strands of analyzed genomes are shown in Table 2. In each of the analyzed genomes, the number of genes located on leading DNA strands is significantly higher than the number of genes located on lagging strands. This phenomenon has been observed several times for many prokaryotic genomes (e.g. Freeman et al. 1998; McLean et al. 1998). We have assumed that if inversion itself is a random process, the probability of inversion of a coding sequence in the direction from the leading strand to the lagging strand should reflect the fraction of genes lying on the leading strand. Inversion in the opposite direction should reflect the fraction of genes lying on the lagging strand. Thus, considering only the orthologs which have changed their positions, the fraction of orthologs which moved from the leading to the lagging strand should reflect the fraction of genes lying on leading strand in each of the two genomes. The same should be true for the lagging strand genes. This is not the case. The numbers of "inverted" orthologs in both pairs of analyzed genomes are shown in Table 2. The null hypothesis that the numbers of genes located on the leading or lagging strands correspond to the numbers of inversions in each direction has to be rejected with a high confidence, with $p < 0.001$ when analyzed by chi square test. In each genome the relative numbers of orthologs inverted from the lagging to the leading strand is higher than the number of orthologs inverted in the opposite direction. These results mean that the substantial fraction of inverted genes from the leading strand to the lagging strand has been killed—such recombinants have been eliminated from populations. Since we have no method for estimating the real fraction of mutated genes eliminated by selection we can only state that the difference between the fraction of the killed genes inverted from the leading to lagging strand and killed genes inverted in the opposite direction corresponds to the difference between fractions of genes nested on the leading strand and the fraction of genes nested on the lagging strand.

In a more formal language we can write:

$$N_{ld}/N_{lg} = Inv_{ld}/Inv_{lg}, \quad (5)$$

$$Inv_{ld} = Inv_{ld\text{-}living} + Inv_{ld\text{-}killed}, \quad (6)$$

Table 2. The number of genes on the leading and lagging strands of the analyzed genomes and the observed number of genes which switched strands

Genome	Number of Genes				Relative Number of Inversions	
	On leading strand (N_{ld})	On lagging strand (N_{lg})	Switched from leading to lagging strand ($Inv_{ld\text{-}living}$)	Switched from lagging to leading strand ($Inv_{lg\text{-}living}$)	From leading to lagging strand ($Inv_{ld\text{-}living}/N_{ld}$)	From lagging to leading strand ($Inv_{lg\text{-}living}/N_{lg}$)
B. burgdorferi	564	286	24	27	0.043	0.094
T. pallidum	678	353	143	136	0.211	0.385
C. pneumoniae	576	476	55	55	0.095	0.116
C. trachomatis	497	396	14	20	0.028	0.051

$$Inv_{lg} = Inv_{lg\text{-}living} + Inv_{lg\text{-}killed}, \quad (7)$$

Since $N_{ld}/N_{lg} > 1$ and $Inv_{ld\text{-}living} \approx Inv_{lg\text{-}living}$, then:

$$Inv_{ld\text{-}killed} > Inv_{lg\text{-}killed} \quad (8)$$

where: N_{ld} and N_{lg}—numbers of genes on the leading and lagging strands, respectively, Inv_{ld} and Inv_{lg}—real (expected) numbers of inversions from the leading to the lagging strand and in the opposite direction respectively, *living*—observed inversions, *killed*—inversions which have been eliminated by selection.

The null hypothesis that the found numbers of inversions in each direction are equal to each other may be accepted with a high confidence (when analyzed by chi square test). The statistically similar numbers of inversions which "have survived" independently of the direction of inversions seem to prove that the analyzed genomes are at steady state. The bias in the distribution of genes between leading and lagging strands is in equilibrium condition in the analyzed genomes.

There are some premises which could explain why recombinants with genes transferred from the leading to the lagging strand are killed more often than recombinants with genes transferred in the opposite direction:

- conserved genes from the leading strand can *accommodate* fewer amino-acid substitutions when affected by a higher mutational pressure after inversion;
- the stronger negative interaction of transcription and replication processes of the highly expressed genes when the genes are transcribed in the direction opposite to the replication fork movement (Brewer 1988; McInerney 1998);
- translation rate of the transcripts could be affected by the accumulation of the nucleotide substitutions in the third codon positions (i.e. Ikemura 1981; Gouy and Gautier 1982; Sharp and Li 1987; Andersson and Kurland 1990; Kanaya et al. 1999) and sometimes in the first codon positions (Gutierrez et al. 1996; Pan et al. 1998) after inversion. Highly expressed genes seem to be more sensitive to such discrimination control through codon usage (it is obvious that the level of degeneracy of genes with high Codon Adaptation Index is lower than average degeneracy of the genetic code).

Conclusions

Our results indicate that the structure of bacterial chromosomes is not random. Inversion of genes with respect to their location on the leading/lagging strand exposes these genes to a higher mutation rate. The more conserved genes are located on the leading strand. Furthermore, they are located on the leading strand because they need to be more conserved and protected from mutations. Transfer of these genes to the lagging strand causes a higher mutation rate and selection leads to their elimination with very high probability. On the other hand, genes located on the lagging strand can accumulate relatively more mutations and still survive. That is why these genes could switch their position with respect to the leading/lagging strand with a lower probability of being killed.

Acknowledgements. Funding sources: This work was supported by The State Committee for Scientific Research, grants number 6 P04A 025-18 and 6 P04A 016 20. P.M. was supported by Foundation for Polish Science.

References

Andersson SG, Kurland CG (1990) Codon preferences in free-living microorganisms. Microbiol Rev 54:198–210

Basic-Zaninovic T, Palombo F, Bignami M, Dogliotti E (1992) Fidelity of replication of the leading and the lagging DNA strands opposite N-methyl-N-nitrosourea-induced DNA damage in human cells. Nucleic Acids Res 20:6543–6548

Bellgard MI, Itoh T, Watanabe H, Imanishi T, Gojobori T (1999) Dynamic evolution of genomes and the concept of genome space. Ann N Y Acad Sci 870:293–300

Blattner FR, Plunkett G, Bloch CA, et al. (1997) The complete genome sequence of *Escherichia coli* K-12. Science 277:1453–1462

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

Dayhoff MO, Schwartz RM, Orcutt BC (1978) A model of evolutionary change in proteins. In: Dayhoff MO (ed) Atlas of protein se-

quence and structure. National Biomedical Research Foundation, Washington, DC, pp 345–352 Vol. 5, Suppl. 3

Felsenstein J (1993) PHYLIP: phylogeny inference package, version 3.5c. Distributed by the author. Department of Genetics, University of Washington, Seattle, WA

Fijałkowska IJ, Jonczyk P, Maliszewska-Tkaczyk M, Bialoskorska M, Schaaper RM (1998) Unequal fidelity of leading strand and lagging strand DNA replication on the *Escherichia coli* chromosome. Proc Natl Acad Sci USA 95:10020–10025

Filipski J (1990) Evolution of DNA sequences. Contributions of mutational bias and selection to the origin of chromosomal compartments. In: Obe G (ed) Advances in mutagenesis. Springer Verlag, Berlin, research 2, pp. 1–54

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

Frank AC, Lobry JR (1999) Asymmetric substitution patterns: a review of possible underlying mutational or selective mechanisms. Gene 238:65–77

Fraser CM, Casjens S, Huang WM, et al. (1997) Genomic sequence of a Lyme disease spirochaete, *Borrelia burgdorferi*. Nature 390:580–586

Fraser CM, Norris SJ, Weinstock GM, et al. (1998) Complete genome sequence of *Treponema pallidum*, the syphilis spirochete. Science 281:375–388

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

French S (1992) Consequences of replication fork movement through transcription units *in vivo.* Science 258:1362–1365

Gouy M, Gautier C (1982) Codon usage in bacteria: correlation with gene expressivity. Nucleic Acids Res 10:7055–7074

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

Gutierrez G, Marquez L, Marin A (1996) Preference for guanosine at first codon position in highly expressed *Escherichia coli* genes. A relationship with translation efficiency. Nucleic Acids Res 24: 2525–2528

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

Itoh T, Takemoto K, Mori H, Gojobori T (1999) Evolutionary instability of operon structures disclosed by sequence comparisons of complete microbial genomes. Mol Biol Evol 16:332–346

Iwaki T, Kawamura A, Ishino Y, Kohno K, Kano Y, Goshima N, Yara M, Furusawa M, Doi H, Imamoto F (1996) Preferential replication-dependent mutagenesis in the lagging DNA strand in *Escherichia coli*. Mol Gen Genet 251:657–664

Kalman S, Mitchell W, Marathe R, Lammel C, Fan J, Hyman RW, Olinger L, Grimwood J, Davis RW, Stephens RS (1999) Comparative genomes of *Chlamydia pneumoniae* and *C. trachomatis.* Nat Genet 21:385–389

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

Karlin S (1999) Bacterial DNA strand compositional asymmetry. Trends Microb 8:305–308

Kolsto AB (1997) Dynamic bacterial genome organization. Mol Microbiol 24:241–248

Koonin EV, Tatusov RL, Galperin MY (1998) Beyond complete genomes: from sequence to structure and function. Curr Opin Struct Biol 8:355–363

Kornberg A, Baker TA (1992) DNA replication. W.H. Freeman, New York

Lafay B, Lloyd AT, McLean MJ, Devine KM, Sharp PM, Wolfe KH (1999) Proteome composition and codon usage in spirochaetes: species-specific and DNA strand-specific mutational biases. Nucleic Acids Res 27:1642–1649

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

Mackiewicz P, Gierlik A, Kowalczuk M, Dudek MR, Cebrat S (1999a) How does replication-associated mutational pressure influence amino acid composition of proteins? Genome Res 9:409–416

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

McLean MJ, Wolfe KH, Devine KM (1998) Base composition skews, replication orientation, and gene orientation in 12 prokaryote genomes. J Mol Evol 47:691–696

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

Mushegian AR, Koonin EV (1996) Gene order is not conserved in bacterial evolution. Trends Genet 12:289–290

Pan A, Dutta C, Das J (1998) Codon usage in highly expressed genes of *Haemophillus influenzae* and *Mycobacterium tuberculosis:* translational selection versus mutational bias. Gene 215:405–413

Roberts JD, Izuta S, Thomas DC, Kunkel TA (1994) Mispair-, site-, and strand-specific error rates during simian virus 40 origin-dependent replication in vitro with excess deoxythymidine triphosphate. J Biol Chem 269:1711–1717

Rocha EP, Danchin A, Viari A (1999) Universal replication biases in bacteria. Mol Microbiol 32:11–16

Salzberg SL, Salzberg AJ, Kerlavage AR, Tomb JF (1998) Skewed oligomers and origins of replication. Gene 217:57–67

Sharp PM, Li WH (1987) The codon adaptation index: a measure of directional synonymous codon usage bias and its potential applications. Nucleic Acids Res 15:1281–1295

Sokal RR, Rohlf FJ (1995) Biometry. Freeman, New York

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

Szczepanik D, Mackiewicz P, Kowalczuk M, Gierlik A, Nowicka A, Dudek MR, Cebrat S (2001) Evolution rates of genes on leading and lagging DNA strands. J Mol Evol 52 (in press)

Tatusov RL, Galperin MY, Natale DA, Koonin EV (2000) The COG database: a tool for genome-scale analysis of protein functions and evolution. Nucleic Acids Res 28:33–36

Tatusov RL, Mushegian AR, Bork P, Brown NP, Hayes WS, Borodovsky M, Rudd KE, Koonin EV (1996) Metabolism and evolution of *Haemophilus influenzae* deduced from a whole-genome comparison with *Escherichia coli.* Curr Biol 6:279–291

Thomas DC, Svoboda DL, Vos JM, Kunkel TA (1996) Strand specificity of mutagenic bypass replication of DNA containing psoralen monoadducts in a human cell extract. Mol Cell Biol 16:2537–2544

Thompson JD, Higgins DG, Gibson TJ (1994) CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, positions-specific gap penalties and weight matrix choice. Nucleic Acids Res 22:4673–4680

Tillier ER, Collins RA (2000a) The contributions of replication orientation, gene direction, and signal sequences to base-composition asymmetries in bacterial genomes. J Mol Evol 50:249–257

Tillier ER, Collins RA (2000b) Replication orientation affects the rate and direction of bacterial gene evolution. J Mol Evol 51:459–463

Trinh TQ, Sinden RR (1991) Preferential DNA secondary structure mutagenesis in the lagging strand of replication in *E. coli.* Nature 352:544–547

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

Watanabe H, Mori H, Itoh T, Gojobori T (1997) Genome plasticity as a paradigm of eubacterial evolution. J Mol Evol 44 (Suppl. 1):S57–S64