J Mol Evol (2001) 52:426–433
DOI: 10.1007/s002390010172

JOURNAL OF MOLECULAR EVOLUTION

Evolution Rates of Genes on Leading and Lagging DNA Strands

Dorota Szczepanik, Paweł Mackiewicz, Maria Kowalczuk, Agnieszka Gierlik, Aleksandra Nowicka, Mirosław R. Dudek, Stanisław Cebrat

Institute of Microbiology, Wrocław University, ul. Przybyszewskiego 63/77, 51-148 Wrocław, Poland

Received: 24 July 2000 / Accepted: 16 January 2001

Abstract. One of the main causes of bacterial chromosome asymmetry is replication-associated mutational pressure. Different rates of nucleotide substitution accumulation on leading and lagging strands implicate qualitative and quantitative differences in the accumulation of mutations in protein coding sequences lying on different DNA strands. We show that the divergence rate of orthologs situated on leading strands is lower than the divergence rate of those situated on lagging strands. The ratio of the mutation accumulation rate for sequences lying on lagging strands to that of sequences lying on leading strands is rather stable and time-independent. The divergence rate of sequences which changed their positions, with respect to the direction of replication fork movement, is not stable—sequences which have recently changed their positions are the most prone to mutation accumulation. This effect may influence estimations of evolutionary distances between species and the topology of phylogenetic trees.

Key words: Orthologs — DNA asymmetry — Mutation pressure — Evolution rate

Introduction

The asymmetry of the DNA molecule is defined as a bias in nucleotide composition of complementary strands, or a deviation from Parity Rule type 2 which says that the Chargaff rules, [A] = [T] and [G] = [C], should be fulfilled not only in the double-stranded DNA molecule, but also in each of the two complementary DNA strands (Sueoka 1995). There are two groups of mechanisms that introduce asymmetry into the DNA molecule, one related to the mutational pressure of the replication and transcription mechanisms, and the second group resulting from selection (see for review: Francino and Ochman 1997; Mrazek and Karlin 1998; Frank and Lobry 1999; Karlin 1999). The divergence of genes observed during phylogeny is a final effect of both groups of mechanisms.

Transcription-associated mutational pressure introduces asymmetry into coding sequences, which could differentiate sense and antisense strands, but this asymmetry is supposed to be independent of the position of a gene on the leading or lagging strand, unless there are differences resulting from different transcription rates (Francino et al. 1996; Francino and Ochman 1997, 1999).

The effect of replication-associated mutational pressure on protein coding sequences is different. It is known that two different mechanisms are involved in replication of the two strands of the DNA molecule, one for the leading strand and the other for the lagging strand (Kornberg and Baker 1992). Thus, it is obvious that different preferences in nucleotide substitutions, and a different frequency of substitutions for the two strands should be expected. Coding sequences themselves are asymmetrical because of their uneven codon composition, implicating preferences in using purines in the sense strand (e.g. Shepherd 1981; Karlin and Burge 1995; Cebrat et al. 1998; McLean et al. 1998). Thus, the positions of protein coding sequences on the leading or lagging

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

strand can determine their susceptibility to mutations. For example, as Frank and Lobry (1999) concluded, if the sense strand is poorer in cytosine, which is the most prone to be substituted by thymine in the leading DNA strand, a gene whose sense strand lies on the leading strand will accumulate fewer such substitutions than a gene whose sense strand lies on the lagging strand (in such a case the antisense strand, rich in cytosine, lies on the leading strand). This not only raises the problem of a higher mutation rate, but also that of a higher rate of defective gene elimination, and/or a higher rate of evolution.

Note that asymmetry between leading and lagging strands does not indicate the absolute values of the mutation rate of the two strands. Measuring asymmetry, we only measure the differences in substitution rates between the two strands. That is why the frequency of substitutions on the two strands problem is still open. Many authors have found that error rates are higher on the lagging strand (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), measuring the mutation rate in the lactose operon incorporated into the *Escherichia coli* chromosome in two opposite directions, have found that mutation rate in the gene inserted in the direction of the leading strand is higher.

As a consequence of asymmetrical processes, compositional bias of DNA strands has been observed in bacterial chromosomes (e.g. Lobry 1996; Freeman et al. 1998; Grigoriev 1998; McLean et al. 1998). It was found that the asymmetry in nucleotide composition changes its sign at the origin and terminus of replication, where DNA strands change their role from leading to lagging and vice versa. The asymmetry is so strong that it is observed even on the level of codons and amino acid composition of proteins (McInerney 1998; Lafay et al. 1999; Mackiewicz et al. 1999a; Rocha et al. 1999).

There are two specific properties of DNA asymmetry which seem to be universal, at least for all eubacterial genomes: the sense strands of protein coding sequences are richer in purines, and the leading strand is richer in guanine.

These properties imply that the substitution rate should be influenced by whether the sense strand of protein coding genes is located on the leading or lagging strand. These properties also imply that the differences in the mutation accumulation rate should be consistent for different genomes. So, it should be expected that some sequences may evolve faster than others, which has been suggested by Radman (1998). Furthermore, since the leading and lagging strands are exposed to different mutational pressures, the inversion of a gene fitted to its position from one strand to the other should lead to a very fast accumulation of mutations in the very first period after inversion, which has been recently suggested by Tillier and Collins (2000). It has been found that mutational patterns vary across mammalian chromosomes and different chromosomal regions evolve at different rates (e.g. Filipski 1988; Wolfe et al. 1989; Boulikas 1992; Eyre-Walker 1992; Holmquist and Filipski 1994; Matassi et al. 1999). Variations in mutation rates across chromosomes correlate with replication timing and could be related to the different efficiency of DNA repair mechanisms. In enterobacterial genomes it was observed that genes located near the origin of replication diverged more slowly than genes located in other parts of chromosome (Sharp et al. 1989; Sharp 1991). Furthermore, the evolutionary rate is smaller for highly expressed genes (Sharp and Li 1987a). In mitochondrial genomes it has been found that mutation rate of their genes differs between regions (Reyes et al. 1998; Koulianos and Crozier 1999).

In this paper we have checked the difference between the divergence rate of genes lying on the leading strand, genes lying on the lagging strand, and genes which changed their positions on the chromosome during phylogeny.

Materials and Methods

Amino acid sequences of orthologs of completely sequenced genomes were extracted from Clusters of Orthologous Groups (COGs), downloaded from ftp://www.ncbi.nlm.nih.gov/pub/COG January 20th, 2000. COGs contain proteins which are supposed to have evolved from one ancestral protein (Koonin et al. 1998; Tatusov et al. 2000). Orthologs are sequences from different species evolved by vertical descent and are usually responsible for the same function in different organisms (Fitch 1970).

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.

The amino acid sequences of each of 2103 COGs were aligned by the CLUSTAL W 1.8 v. program (Thompson et al. 1994). To estimate evolutionary distances, the pairwise distances between sequences in each COG were calculated with the CLUSTAL W 1.8 v. program, using Kimura correction for multiple substitutions (Kimura 1983), with improvements for very diverged sequences made by the authors of CLUSTAL W 1.8 v., and also 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).

Analyses have been performed with 12 645 orthologs derived from 11 eubacterial genomes showing evident compositional asymmetry between leading and lagging strands. For each pair of organisms, orthologs were classified into three groups according to their strand location: sequences lying on leading strands (in both compared genomes), sequences lying on lagging strands, and sequences which changed their position. The mean value of the evolutionary distances was calculated for each group. Nonparametric analysis by Mann–Whitney U, Kolmogorov–Smirnov and ANOVA Kruskal–Wallis tests (Sokal and Rohlf 1995) were carried out to assess statistical significance between these groups.

The neighbor-joining method was used to construct trees for the three groups of orthologs (with the program NEIGHBOR from the PHYLLIP package).

Boundaries between the leading and lagging strands (position of origin and terminus of replication) and decisions concerning the loca-

Table 1. The mean evolutionary distances between orthologs from leading and lagging strands

	Bb	Bs	Cp	Ct	Ec	Hi	Hpy	Hpj	Mt	Rp	Tp
Bb		**2.54**	**2.87**	**2.60**	**2.42**	**1.96**	**2.16**	**2.24**	**2.29**	1.98	3.21
Bs	**3.37**		**3.12**	**2.93**	**3.32**	**3.06**	**2.99**	**3.06**	**3.20**	**3.24**	**2.66**
Cp	**2.92**	**3.96**		**0.90**	**3.37**	**2.78**	**2.24**	**2.60**	**3.72**	2.35	3.15
Ct	**2.98**	**4.12**	**0.97**		**3.34**	**2.54**	**2.51**	**2.54**	**3.09**	2.20	2.75
Ec	**3.11**	**3.89**	**3.31**	**3.33**		**2.53**	**3.09**	**3.24**	**3.55**	**3.30**	**2.76**
Hi	**3.15**	**3.46**	**3.12**	**2.82**	**3.04**		**2.67**	**2.83**	**3.33**	**2.53**	**2.44**
Hpy	**2.96**	**4.10**	**2.75**	**2.81**	**3.98**	**3.54**		0.79	**3.54**	**2.52**	**2.51**
Hpj	**2.94**	**4.16**	**2.76**	**2.80**	**4.03**	**3.45**	0.92		**3.74**	**2.65**	**2.62**
Mt	**5.09**	**3.79**	**4.07**	**3.90**	**3.96**	**4.44**	**4.23**	**4.24**		**3.32**	3.06
Rp	2.21	**4.09**	2.25	2.10	**3.71**	**2.59**	**2.84**	**2.87**	**3.51**		2.50
Tp	3.23	**3.07**	3.31	**3.05**	**2.84**	**3.23**	**2.72**	**2.70**	2.79	2.74	

The mean value of the evolutionary distances for protein sequences based on the Dayhoff PAM matrix model, calculated for each pair of genomes, separately for orthologs lying on the leading strand (upper right triangle), and those on the lagging strand (lower left triangle). the values are the mean of amino acid substitutions per site between two genomes. Pairs of values which are statistically different ($P < 0.05$) for the two strands are in **bold.** Abbreviations: Bb: *Borrelia burgdorferi,* Bs: *Bacillus subtilis,* Cp: **Chlamydia pneumoniae,** Ct: *Chlamydia trachomatis,* Ec: *Escherichia coli,* Hi: *Haemophilus influenzae,* Hpy: *Helicobacter pylori 26695,* Hpj: *Helicobacter pylori J99,* Mt: *Mycobacterium tuberculosis,* Rp: *Rickettsia prowazekii,* Tp: *Treponema pallidum.*

Table 2. The mean evolutionary distances between orthologs which changed their location

	Bb	Bs	Cp	Ct	Ec	Hi	Hpy	Hpj	Mt	Rp	Tp
Bb		2.89	3.20	2.94	2.98	2.74	2.69	2.77	3.46	2.51	3.92
Bs	0.000		3.39	3.26	3.66	3.22	3.52	3.56	3.52	3.48	2.96
Cp	0.022	0.000		3.04	3.55	2.98	2.84	3.00	3.91	2.45	3.50
Ct	0.004	0.000	0.000		3.28	2.69	2.88	2.93	3.78	2.47	3.10
Ec	0.001	0.000	0.000	0.000		2.96	3.57	3.57	3.80	3.55	2.90
Hi	0.000	0.000	0.001	0.001	0.000		2.83	2.89	3.84	2.90	3.10
Hpy	0.000	0.000	0.000	0.000	0.000	0.000		2.77	4.10	2.95	3.11
Hpj	0.000	0.000	0.000	0.000	0.000	0.000	0.000		4.34	2.89	2.88
Mt	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000		3.63	3.58
Rp	0.008	0.000	0.212	0.016	0.000	0.013	0.006	0.043	0.000		2.40
Tp	0.000	0.006	0.110	0.042	0.002	0.000	0.000	0.033	0.000	0.331	

Distance matrix measured for orthologs which changed their location from the leading to lagging strand within a pair of genomes (upper right triangle) and significance level between three groups of orthologs (lying on the leading strand, lying on the lagging strand and the ones which changed strand) analysed by the AVOVA Kruskal-Wallis test (lower left triangle). Abbreviations are as in Table 1.

tion of genes on one of these strands were set on the basis of experimental results (indicated in data bases listed below), or on the basis of DNA walk results that described the nucleotide compositional bias of DNA strands (Mackiewicz et al., 1999a, see also: http://smorfland.microb.uni.wroc.pl).

Prokaryotic genomic sequences were downloaded from ftp://www.ncbi.nlm.nih.gov: *Borrelia burgdorferi* (Fraser et al. 1997), *Bacillus subtilis* (Kunst et al. 1997), *Chlamydia trachomatis* (Stephens et al. 1998), *Chlamydia pneumoniae* (Kalman et al. 1999), *Escherichia coli* (Blattner et al. 1997), *Haemophilus influenzae* (Fleischmann et al. 1995), *Helicobacter pylori 26695* (Tomb et al. 1997), *Helicobacter pylori J99* (Alm et al. 1999), *Mycobacterium tuberculosis* (Cole et al. 1998), *Rickettsia prowazekii* (Andersson et al. 1998), *Treponema pallidum* (Fraser et al. 1998).

Results

For each pair of genomes, the mean value of evolutionary distances for protein sequences, based on the Dayhoff PAM matrix model (Dayhoff et al. 1978), was calculated separately for leading strand orthologs (Table 1, upper right triangle), and lagging strand orthologs (Table 1, lower left triangle). The values are the mean number of amino acid substitutions between two genomes, per site. Value pairs which are statistically different ($P < 0.05$) for the two strands are in bold.

In almost all cases, distances between genomes measured by divergence of orthologs from the lagging strand are larger than distances counted on the basis of the leading strand orthologs. Almost all the differences between these distances (measured for orthologs from leading and lagging strand) are statistically significant. In some cases, distances for the leading strand orthologs are greater than distances for the lagging strand orthologs. Nevertheless, in all such cases the differences are very small or are not statistically significant.

The distance matrix measured for orthologs which changed their location from the leading to lagging strand or vice versa within a pair of genomes has been shown in Table 2 (upper right triangle). For closely related pairs of genomes (especially *Chlamydia* species, *Helicobacter*

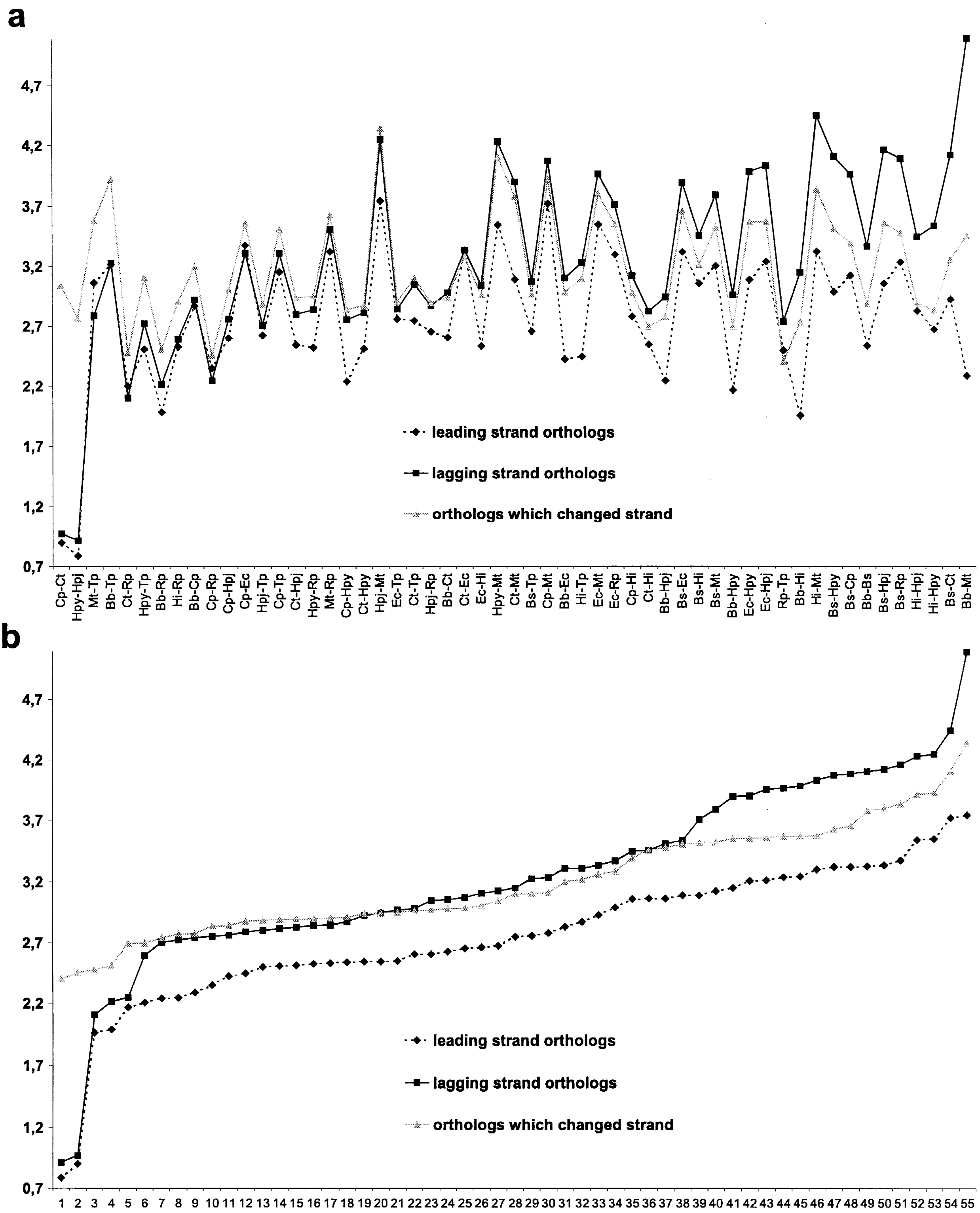

Fig. 1. The mean value of the evolutionary distances for protein sequences based on the Dayhoff PAM matrix model, calculated for each pair of genomes, separately for leading strand orthologs, lagging strand orthologs, and orthologs which switched strands. The values are the mean numbers of amino acid substitutions per site between two genomes: **(A)** data have been arranged according to the difference in distance for orthologs from the lagging strand, orthologs which switched the strand in the given pair of genomes; **(B)** data have been arranged by groups of orthologs. Abbreviations are as in Table 1.

pylori strains and the spirochaete bacteria, *B. burgdorferi,* and *T. pallidum*), these distances are larger than distances measured for lagging and leading strand orthologs. For more distant pairs of genomes, the distances measured for orthologs which changed their locations decrease relatively and stay between values for orthologs from the leading and lagging strands. This relation between the mutation accumulation rate and the location of

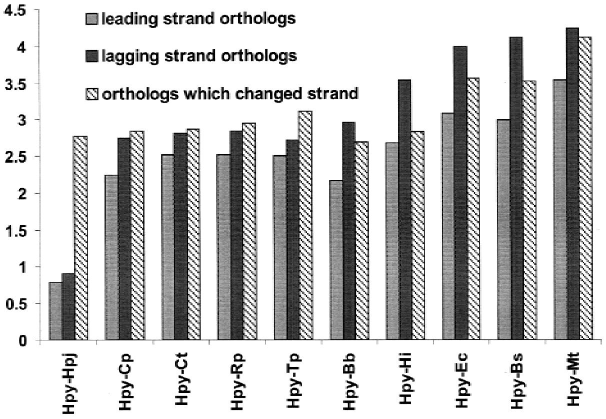

Fig. 2. Distances (mean number of amino acid substitutions per site, calculated using the Dayhoff PAM matrix model) between *H. pylori 26695* genome and the rest of the analysed genomes for three groups of orthologs. Abbreviations are as in Table 1.

orthologs is seen in Fig. 1*a*. In Fig. 1*b* the data have been arranged in separate ortholog group rankings. In this figure it is clear that the dynamics of mutation accumulation for orthologs which have changed their positions is different than for two other groups of orthologs. While the ratio of the distance measured for groups of genes lying on the lagging strand to that for the genes on the leading strand is stable (correlation coefficient = 0.78), the ratio of the distances measured for genes which changed their position, to distances measured for the other groups changes in evolutionary time. It is also seen when one genome is compared with the rest of the analysed genomes (Fig. 2).

In almost all cases, the three groups of analysed orthologs create sets which significantly differ, statistically, in the rate of evolution when analysed by the AVOVA Kruskal–Wallis test (Table 2, lower left triangle).

To show the influence of different gene evolution rates on the topology of the phylogenetic tree, we have constructed such trees for eight genomes, based on genes located on the leading strand, the lagging strand, and on genes which switched their strand (Fig. 3). Note that only orthologs from the leading strand can be used to construct a proper phylogenetic tree. In this case it is possible to find groups of orthologs representing all genomes used for analysis. We have not found such groups of orthologs on the lagging strand and representing all genomes. It is obvious that genes which switched their positions could be used during the construction of such a tree only in pairs of genomes. That is why we prepared all three trees based on all the orthologs found in proper relations, with respect to the strand in each pair of compared genomes. As it can be seen, the topology of the trees for orthologs from leading and lagging strands is not very different, besides the longer distances for the lagging strands. Nevertheless, in the tree based on orthologs which switched their strand, not only were absolute distances changed, but also the topology of the tree, especially for closely related genomes. Thus, it seems important to consider the position of genes when they are chosen for estimating the phylogenetic relations between organisms.

Discussion

We have found that the values of divergence in the group of leading strand orthologs is significantly lower, statistically, than that of lagging orthologs. There are two possible explanations of this phenomenon: the absolute mutation level in genes, whose transcription direction is in agreement with the direction of the movement of replication fork, is lower, or the selective constraints on gene sequences lying in this direction are higher.

If we assume that the second explanation is correct, it means that genes lying on the leading strand are more conservative. Nevertheless, it seems plausible to accept the hypothesis that both mechanisms are important in establishing the final level of accumulated mutations. Thus, one should accept that selection also plays a crucial role in establishing the mutation rate of more conservative genes, on a lower level, by leaving them on the leading strand, while genes which should evolve faster are located on the lagging strand. Sharp and Li (1987a) found that highly expressed genes have smaller evolutionary rates. That seems to be true, since an overwhelming part of genes coding for ribosomal proteins are located on the leading strand in every genome thus far sequenced (e.g. McLean et al. 1998). Their operons and several others are found to be well conserved during evolution across genomes (Kunst et al. 1997; Watanabe et al. 1997; Itoh et al. 1999). Location of genes with higher expressivity on the strand where the mutational pressure is lower seems to be in agreement with the implication of the second law of thermodynamics: More frequent reading of the information from the source introduces more noise to this source. If the genes with high expressivity were not located at a position with lower mutation rate, they would have to accumulate mutations with higher rate than observed. On the other hand, it is reasonable to assume that genes fit to the strand they are located on. “Fit” means that their nucleotide compositions and their own asymmetry guarantee a lower mutation rate in the position they are in. The time-dependent divergence rate of genes which have changed their location confirms this hypothesis. The average divergence in this group of orthologs is higher than for the lagging strand in closely related genomes, which means that gene inversion immediately causes a higher mutation rate and, if the gene survives it becomes “better fitted” to the new location. For example, if the gene is located in such a way that its strand richer in cytosine is located on the strand where cytosine preferentially undergoes transition to thymine, after the elimination of some cytosine resi-

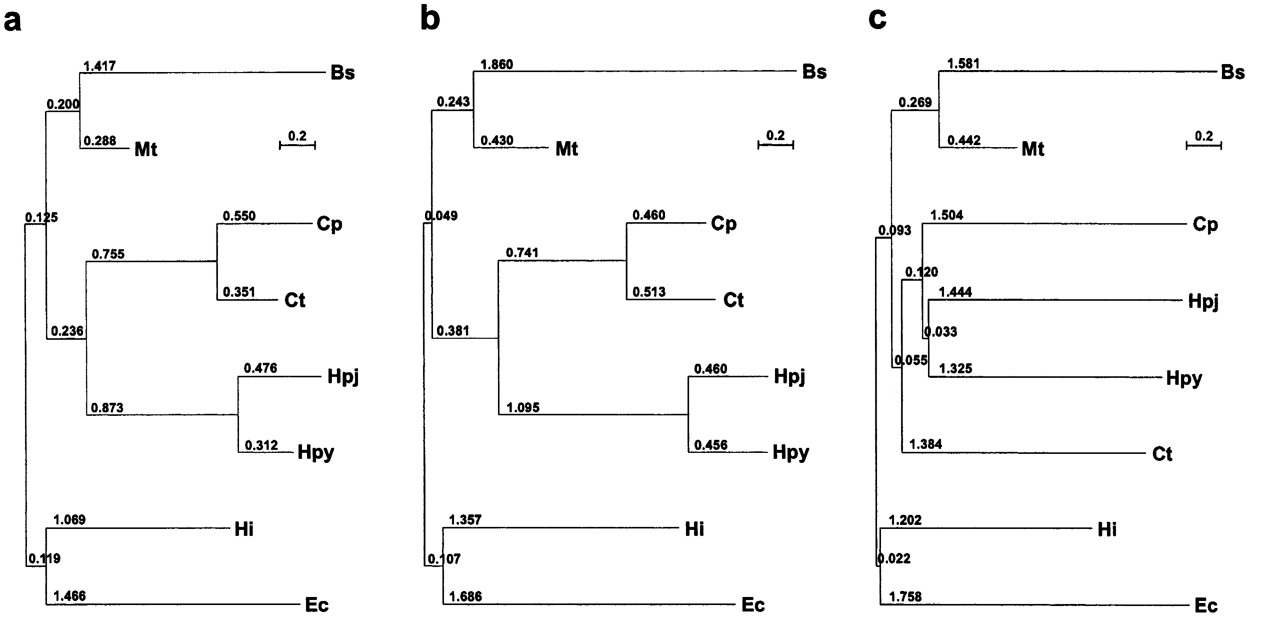

Fig. 3. Neighbor-joining trees constructed for orthologs: **(A)** from the leading strand; **(B)** from the lagging strand; **(C)** which changed strand. The scale bar corresponds to 20 amino acid substitutions per 100 positions. The numbers indicate the length of the branches. Abbreviations are as in Table 1.

dues (if it is not lethal) the overall mutation rate would diminish with time. We have observed such susceptibility of genes translocated with inversion to mutations in computer simulations (results not shown). This phenomenon also explains the results of Fijałkowska et al. (1998), who found that the gene *lacZ* of lactose operon cumulates more mutations if it is on the leading strand which, at first, seems to disagree with our results. But the usual location of lactose operon in the *E. coli* genome is on the lagging strand. Thus, the effect observed by these authors is an effect of inversion, not the effect of a higher mutation rate on the leading strand. The effect of a higher mutation rate in inverted sequences diminishes very fast with time measured in evolutionary scale (Fig. 1*b*). Nevertheless, under laboratory conditions the effect should be very strong since only the very first replication cycles are observed when the translocated sequence is in strong disequilibrium and is very prone to substitutions.

The PAM matrices (Dayhoff et al. 1978) used for distance estimations have built-in selection parameters. One can argue that the observed differences result from using different weights for amino acid substitutions for different sets of proteins and, in fact, selection is the main force positioning specific genes on specific strands. That is why we have repeated our calculations of divergence using the Kimura algorithm (Kimura 1983), assuming the neutrality of substitutions. Both methods give qualitatively the same results: The correlation coefficient between distances counted on the base of PAM matrix and Kimura algorithm was 0.97, though the Kimura algorithm gave slightly lower values of divergence.

The general conclusion that the divergence rate is higher for the lagging strand is further supported by the fact that the fraction of proteins in COGs coded by genes lying on lagging strands of genomes is underrepresented. The ratio between the number of leading strand genes to the number of lagging strand genes is higher in the genome data bases than in COGs. The difference is statistically significant. It could happen if some orthologs are not recognised as homologues because the divergence is too high. If the lagging strand genes are underrepresented it is because they have been preferentially omitted during the construction of COGs.

There is still an unsolved problem: What is the main cause of the observed asymmetrical evolution rates of the leading and lagging strands? Is it replication-associated mutational pressure, or the selection leaving conservative genes on the leading strand? Recent studies of Mackiewicz et al. (1999a) have allowed the recognition of the effect of replication-associated mutational pressure from other effects introducing asymmetry into DNA, connected with transcription and other coding functions. Analysing the pure effect of replication they found that the specific asymmetry introduced by replication-associated mutational pressure into intergenic sequences could be also recognised in each position of codons, though of different magnitude. It means that genes lying on a given strand show some trends in nucleotide composition consistent with the trend introduced by substitutions associated with replication, even in the most conservative second positions in codons. Lafay et al. (1999) have found that some orthologs which have changed their positions in the *T. pallidum* and *B. burgdorferi* genomes have accumulated mutations which have assimilated them to the new positions, and now show codon and amino acid compositions typical of their current location. In fact, protein coding genes in the *B. burgdorferi* genome form two distinct, nonoverlapping sets which show the distinct effect characteristic for replication-associated mutational pressure of each of the two DNA strands (McInerney 1998; Lafay et al. 1999; Mackiewicz et al. 1999b).

The translation efficiency can be influenced by the codon usage (i.e. Ikemura 1981; Gouy and Gautier 1982; Sharp and Li 1987b; Andersson and Kurland 1990; Kanaya et al. 1999), amino acid composition (Lobry and Gautier 1994) and base composition of the first positions of codons (Gutierrez et al. 1996; Pan et al. 1998). Since the replication-associated mutation pressure differentially affects the aforementioned properties of protein coding sequences according to strand location, the translocation of a gene between strands may change its translational efficiency. Nevertheless, the conclusive results would be supplied by the analyses of iso-acceptor t-RNA abundance in the cell, and its influence on the translation rate of genes which accumulated mutations after the inversion.

One could argue that it is transcription-associated mutational pressure which is mainly responsible for the introduction of substitutions into coding sequences (Francino et al. 1996; Francino and Ochman 1997, 1999). If it is true, there should be no characteristic time-dependent changes in the ratio of observed substitutions after translocation with inversion. Even if we assume that some other mechanisms like lateral transfer or reticulate evolution influence the rate of mutation accumulation, we should accept the hypothesis that these mechanisms affect the two DNA strands asymmetrically. We do not see any premises which would allow us to accept such a solution.

In conclusion, we state that the level of mutations introduced during replication is higher for the lagging strand. This leads to the higher level of accumulation of mutations by lagging strand genes and to the asymmetry specific for that introduced by replication-associated mutations. As a result, the selection tends to fix the genes which should be more conservative on the leading strand, enforcing the effect of an asymmetrical gene evolution rate. Both effects are seen even in the asymmetrical amino acid composition of proteins coded by leading and lagging strands. Transposition of an inverted gene (with respect to the leading/lagging strand) has an immediate, strong mutagenic effect. If the gene survives, it can be fixed at the new position, and will eventually change its nucleotide composition to better fit the mutational pressure of the new position. The different mutational pressure of two DNA strands has another evolutionary consequence: It groups genes for slower and faster evolution which may play a big role in adaptation, and may contribute to higher fitness in the fast changing environment.

Note Added in Proof

Recently E.R.M. Tillier and R.A. Collins [*J Mol Evol* (2000) 51:459–463] have shown higher divergence rate for *Chlamydia* genes which switched their positions between the leading and the lagging DNA strands.

Acknowledgements. This work was supported by The State Committee for Scientific Research, grant number 6 P04A 025-18.

References

Alm RA, Ling L-SL, Moir DT, et al. (1999) Genomic-sequence comparison of two unrelated isolates of the human gastric pathogen *Helicobacter pylori.* Nature 397:176–180

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

Andersson SG, Zomorodipour A, Andersson JO, Sicheritz-Ponten T, Alsmark UC, Podowski RM, Naslund AK, Eriksson AS, Winkler HH, Kurland CG (1998) The genome sequence of *Rickettsia prowazekii* and the origin of mitochondria. Nature 396:133–140

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

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

Boulikas T (1992) The evolutionary consequences of nonrandom damage and repair of chromatin domains. J Mol Evol 35:156–180

Cebrat S, Dudek MR, Mackiewicz P (1998) Sequence asymmetry as a parameter indicating coding sequence in *Saccharomyces cerevisiae* genome. Theory Bioscienc 117:78–89

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

Dayhoff MO, Schwartz RM, Orcutt BC (1978) A model of evolutionary change in proteins. In: Dayhoff MO (ed) Atlas of protein sequence and structure Vol. 5, Suppl. 3. National Biomedical Research Foundation, Washington, DC pp 345–352

Eyre-Walker A (1992) The role of DNA replication and isochores in generating mutation and silent substitution rate variance in mammals. Genet Res 60:61–67

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

Fijalkowska 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 (1988) Why the rate of silent codon substitutions is variable within a vertebrate's genome. J Theor Biol 134:159–164

Fitch WM (1970) Distinguishing homologous from analogous proteins. Syst Zool 19:99–113

Fleischmann RD, Adams MD, White O, et al. (1995) Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd. Science 269:496–512

Francino MP, Chao L, Riley MA, Ochman H (1996) Asymmetries generated by transcription-coupled repair in enterobacterial genes. Science 272:107–109

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

Francino MP, Ochman H (1999) A comparative genomics approach to DNA asymmetry. Ann NY Acad Sci 870:428–431

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

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

Holmquist GP, Filipski J (1994) Organization of mutations along the genome: a prime determinant of genome evolution. Trends Ecol Evol 9:65–69

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

Karlin S, Burge C (1995) Dinucleotide relative abundance extremes: a genomic signature. Trends Genet 11:283–290

Kimura M (1983) The neutral Theory of Molecular Evolution. Cambridge University Press, Cambridge

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. WH Freeman and Co., New York

Koulianos S, Crozier RH (1999) Current intraspecific dynamics of sequence evolution differs from long-term trends and can account for the AT-richness of honeybee mitochondrial DNA. J Mol Evol 49:44–48

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

Lafay B, Lloyd AT, McLean MJ, Devine KM, Sharp PM, Wolfe KH (1999) Proteome composition and codon usage in spirochaete: 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

Lobry JR, Gautier C (1994) Hydrophobicity, expressivity and aromaticity are the major trends of amino-acid usage in 999 *Escherichia coli* chromosome-encoded genes. Nucleic Acids Res 22:3174–3180

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

Mackiewicz P, Gierlik A, Kowalczuk M, Szczepanik D, Dudek MR, Cebrat S (1999b) Mechanisms generating long-range correlation in nucleocide composition of the *Borrelia burgdorferi.* Physica A 273:103–115

Matassi G, Sharp PM, Gautier C (1999) Chromosomal location effects on gene sequence evolution in mammals. Curr Biol 9:786–791

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

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

Radman M (1998) DNA replication: one strand may be more equal. Proc Natl Acad Sci USA 95:9718–9719

Reyes A, Gissi C, Pesole G, Saccone C (1998) Asymmetrical directional mutation pressure in the mitochondrial genome of mammals. Mol Biol Evol 15:957–966

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

Sharp PM (1991) Determinants of DNA sequence divergence between *Escherichia coli* and *Salmonella typhimurium:* codon usage map position, and concerted evolution. J Mol Evol 33:23–33

Sharp PM (1991) Determinants of DNA sequence divergence between *Escherichia coli* and *Salmonella typhimurium:* codon usage map position, and concerted evolution. J Mol Evol 33:23–33

Sharp PM, Li WH (1987a) The rate of synonymous substitution in enterobacterial genes is inversely related to codon usage bias. Mol Biol Evol 4:222–230

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

Shepherd JC (1981) Method to determine the reading frame of a protein from the purine/pyrimidine genome sequence and its possible evolutionary justification. Proc Natl Acad Sci USA 78:1596–1600

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

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

Sueoka N (1995) Intrastrand parity rules of DNA base composition and usage biases of synonymous codons. J Mol Evol 40:318–325

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

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

Tomb JF, White O, Kerlavage AR, et al. (1997) The complete genome sequence of the gastric pathogen *Helicobacter pylori. Nature* 388: 539–547

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

Wolfe KH, Sharp PM, Li WH (1989) Mutation rates differ among regions of the mammalian genome. Nature 337:283–285