J. Appl. Genet. 44(4), 2003, pp. 561-584

High divergence rate of sequences located on different DNA strands in closely related bacterial genomes

Paweł MACKIEWICZ[1], Dorota MACKIEWICZ[1], Maria KOWALCZUK[1], Małgorzata DUDKIEWICZ[1], Mirosław R. DUDEK[2], Stanisław CEBRAT[1]

[1]Institute of Genetics and Microbiology, Wrocław University, Wrocław, Poland
[2]Institute of Physics, University of Zielona Góra, Zielona Góra, Poland

Abstract. One of the common features of bacterial genomes is a strong compositional asymmetry between differently replicating DNA strands (leading and lagging). The main cause of the observed bias is the mutational pressure associated with replication. This suggests that genes translocated between differently replicating DNA strands are subjected to a higher mutational pressure, which may influence their composition and divergence rate. Analyses of groups of completely sequenced bacterial genomes have revealed that the highest divergence rate is observed for the DNA sequences that in closely related genomes are located on different DNA strands in respect to their role in replication. Paradoxically, for this group of sequences the absolute values of divergence rate are higher for closely related species than for more diverged ones. Since this effect concerns only the specific group of orthologs, there must be a specific mechanism introducing bias into the structure of chromosome by enriching the set of homologs in trans position in newly diverged species in relatively highly diverged sequences. These highly diverged sequences may be of varied nature: (1) paralogs or other fast-evolving genes under weak selection; or (2) pseudogenes that will probably be eliminated from the genome during further evolution; or (3) genes whose history after divergence is longer than the history of the genomes in which they are found. The use of these highly diverged sequences for phylogenetic analyses may influence the topology and branch length of phylogenetic trees. The changing mutational pressure may contribute to arising of genes with new functions as well.

Key words: divergence, DNA asymmetry, lagging strand, leading strand, mutation pressure.

Received: May 19, 2003. Accepted: July 25, 2003.
Correspondence: S. CEBRAT, Institute of Genetics and Microbiology, Wrocław University, ul. Przybyszewskiego 63/77, 51-148 Wrocław, Poland, email: cebrat@microb.uni.wroc.pl

Introduction

The fraction of positions that are different in the compared sequences is one of the measures of their divergence. The fraction of nucleotides that have been substituted depends directly on the mutational pressure and reciprocally on the selection pressure, and this fraction is supposed to be positively correlated with the time elapsed from the moment of divergence of the compared sequences. That is why the number of substitutions in DNA sequences forms the basis for estimating the phylogenetic distances and evolutionary relationships between compared sequences.

One of the problems of molecular phylogenetic methods is that the accumulated nucleotide substitutions, directly observed during the comparison of sequences, are only a fraction of the substitutions that have happened since the time of divergence. Some substitutions have been eliminated directly by selection forces, some as mutations "accompanying" others selected directly against. The efficacy of selection against a substitution depends on the function of the sequence. It is assumed that mutations in very important housekeeping genes are strongly selected against and that is why such genes are highly conserved (SHARP, LI, 1987). For a long time, to simplify calculations, it has been assumed that the mutational pressure is rather stable or its changes do not disturb significantly the output of the observed fraction of accumulated mutations. However, the discovery of the asymmetric structure of prokaryotic chromosomes has changed this view (e.g. LOBRY 1996, BLATTNER et al. 1997, KUNST et al. 1997, FRASER et al. 1997, 1998, ANDERSSON, ANDERSSON 1999, FREEMAN et al. 1998, GRIGORIEV 1998, MCLEAN et al. 1998, MACKIEWICZ et al. 1999a, ROCHA et al. 1999, TILLIER, COLLINS 2000b, LOPEZ, PHILIPPE 2001). There are two different mutational pressures, introducing substitutions with different preferences into leading and lagging DNA strands (for review see: FRANCINO, OCHMAN 1997, MRAZEK, KARLIN 1998, FRANK, LOBRY 1999, KOWALCZUK et al. 2001a). These different mutational pressures are caused by different replication modes of the leading and lagging strands, which is connected with architectural asymmetry of the replication fork, different processivity (tendency to remain on a single template) of enzyme complexes, different error rates and effectiveness of repair systems on the DNA strands. The cytosine deamination theory (FRANK, LOBRY 1999) assumes that stretches of the template for the newly synthesized lagging strand are temporarily single-stranded. In this state the template is more exposed to damage and mutations, of which the most frequent is deamination of cytosine and its methylated derivative 5-methylocytosine to uracil and thymine, respectively. This leads in consequence to C $\rightarrow$ T transition on the leading strand. Another, less common, A $\rightarrow$ G transition resulting from deamination of adenine to hypoxantine may be connected with asymmetric mutational pressure as well. Different and asymmetric mutational pressures are responsible for different nucleotide compositions of the two DNA strands and different sensitivity of coding

sequences to mutational pressure. Generally, it seems that coding sequences are adjusted to their location on chromosomes and show some asymmetry resulting from the mutational pressure even at the level of codons and amino acids (PERRIERE et al. 1996, MCINERNEY 1998, LAFAY et al. 1999, MACKIEWICZ et al. 1999b, ROCHA et al. 1999, ROMERO et al. 2000). The parameters of the mutational pressure are such that the fraction of a nucleotide in the DNA sequence fully accommodated to the pressure is linearly correlated with the turnover rate of this very nucleotide, measured by its half-time of substitution. Due to this, a sequence that stays longer on the same DNA strand accumulates less mutations per time unit (KOWALCZUK et al. 2001b). Because of different mutational pressures acting on leading and lagging strands, translocation of a sequence already accommodated to its location should result in its higher mutation rate, which was actually observed in closely related genomes (TILLIER, COLLINS 2000a, MACKIEWICZ et al. 2001, ROCHA, DANCHIN 2001, SZCZEPANIK et al. 2001). This is not the only mechanism introducing differences in the rate of accumulation of substitutions and consequently in the rate of divergence. Another source of such differences is selection. In all sequenced genomes, a specific redundancy of genetic information is observed. It is seen in the specific distribution of gene families and paralogs (HUYNEN, van NIMWEGEN 1998, SŁONIMSKI et al. 1998, QIAN et al. 2001), which are homologous sequences existing in the same genome (FITCH 1970). Some of them execute the same or similar function and even complement each other, so they could enhance the viability of organisms being under the mutational pressure (CEBRAT, STAUFFER 2002). Some paralogs fulfill different functions, and some of them are probably inactive pseudogenes that have lost their primary function after duplication. If this is true, the last group should diverge very fast, accumulating all substitutions introduced into their sequences by mutational pressure (LI et al. 1981, GOJOBORI et al. 1982). Moreover, it was found that paralogs evolve under a purifying selection in prokaryotic as well as in eukaryotic genomes (LYNCH, CONERY 2000, KONDRASHOV et al. 2002). Pseudogenes have been identified in many bacterial (DELORME et al. 1993, FRASER et al. 1997, ANDERSSON et al. 1998, ANDERSSON, ANDERSSON 1999, 2001, COLE et al. 2001, MIRA et al. 2001, OGATA et al. 2001, PARKHILL et al. 2001, HOMMA et al. 2002) and eukaryotic genomes (for review see: MIGHELL et al. 2000, HARRISON, GERSTEIN 2002). In this paper we show that homologous sequences located on different (leading/lagging) DNA strands in closely related genomes have accumulated more substitutions than the sequences that have stayed on the same DNA strand. This suggests that a lot of these sequences may be paralogs or sequences evolving under weaker selection, or they are pseudogenes generated from genes subjected to a high mutational pressure after inversion, or their history after divergence is longer than the history of the compared genomes. Moreover, these sequences can significantly disturb the phylogenetic analyses of taxa, giving a paradoxical effect that divergence rate between closely related genomes is higher than between more distant ones.

Material and methods

Data for analysis

Analyses have been done on three sets of bacterial genomes showing evident compositional asymmetry between the leading and lagging DNA strands:

– 11 pairs of closely-related genomes in the COG (Clusters of Orthologous Groups) database:

Escherichia coli K12-MG1655 (EcK) – *E. coli* O157:H7 EDL933 (EcE); *Helicobacter pylori* 26695 (Hp) – *H. pylori* J99 (HpJ); *Neisseria meningitidis* MC58 (NmM) – *N. meningitidis* Z2491 (NmZ); *Bacillus halodurans* (Bh) – *B. subtilis* (Bs); *Chlamydia pneumoniae* (Cp) – *C. trachomatis* (Ct); *Mycobacterium leprae* (Ml) – *M. tuberculosis* (Mt); *Pyrococcus abyssi* (Pab) – *P. horikoshii* (Ph); *Borrelia burgdorfer*i (Bb) – *Treponema pallidum* (Tp); *Caulobacter crescentus* (Cc) – *Mesorhizobium loti* (Mlt); *Haemophilus influenzae* (Hi) – *Pasteurella multocida* (Pm); *Lactococcus lactis* (Ll) – *Streptococcus pyogenes* (Sp);

– 7 genomes belonging to the γ-subdivision of Proteobacteria, compared with one another:

E. coli K12-MG1655 (EcK), *E. coli* O157:H7 EDL933 (EcE), *H. influenzae* (Hi), *P. multocida* (Pm), *Pseudomonas aeruginosa* (Pa), *Vibrio cholerae* (Vc), *Xylella fastidiosa* (Xf);

– 13 genomes compared with *E. coli* O157:H7 EDL933:

B. subtilis, *Campylobacter jejuni*, *C. pneumoniae*, *E. coli* K12-MG1655, *H. pylori* 26695, *M. tuberculosis*, *N. meningitidis* MC58, *P. multocida*, *P. aeruginosa*, *Ricketsia prowazekii*, *T. pallidum*, *V. cholerae*, *X. fastidiosa*.

Moreover, from the 7 genomes of the γ-Proteobacteria group, the 7 sets of 1521 orthologs present in all the genomes, being the "best hits" for *E. coli* EDL933 sequences (the closest orthologs), were withdrawn to construct phylogenetic trees.

Prokaryotic genomic sequences and gene annotations have been downloaded from the Genbank (ftp://www.ncbi.nlm.nih.gov). Boundaries between leading and lagging strands (positions of origins and termini of replication) and decisions concerning the location of genes on one of these strands, were set on the basis of experimental results or on the basis of the results of DNA walks describing a nucleotide compositional bias of differently replicating DNA strands (MACKIEWICZ et al. 1999a, 1999b; see also: http://smorfland.microb.uni.wroc.pl). This method differs from the originally proposed method (LOBRY 1996, GRIGORIEV 1998) in that it cumulates for a given chromosome region local deviations of a parameter of asymmetry (e.g. [G]-[C]) from the average value specific for the whole chromosome. This method eliminates the global compositional trend of the whole chromosome and smoothes random fluctuation. The main switch points of a DNA walk are presumed to be the origin and terminus of replication.

Amino acid sequences of orthologous proteins encoded by the analyzed genomes were extracted from the COG database downloaded from ftp://www.ncbi.nlm.nih.gov/pub/COG in September 2001. COGs contain protein sequences that are supposed to have evolved from one ancestral protein (KOONIN et al. 1998, TATUSOV et al. 2001). Orthologs are sequences from different species that evolved by vertical descent and are usually responsible for the same function in different organisms (FITCH 1970). In the construction of COGs their 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 COG were aligned by the CLUSTAL W 1.8 v. software (THOMPSON et al. 1994). Pairwise evolutionary distances (expressed by the mean number of amino acid substitutions per site) between sequences of each COG were calculated by using the Dayhoff PAM model (DAYHOFF et al. 1978) as implemented in the PROTDIST program of the PHYLIP 3.5c package (FELSENSTEIN 1993). The sequences were analyzed in two ways. In the first approach, the best matches for each ortholog (the closest orthologs) were chosen. In the second approach (assuming a more restrictive definition of orthology), only bidirectional best matches were analyzed.

For each pair of genomes, the orthologs were divided into three groups according to their strand location: (i) pairs of orthologs lying on leading strands in both compared genomes, (ii) pairs of orthologs lying on lagging strands, and (iii) pairs of orthologs of which one is lying on the leading and the other one on the lagging strand (which will be referred to as trans-orthologs). For each of the three groups, the mean values of the evolutionary distances were calculated. Nonparametric analyses by Mann-Whitney U and Kolmogorov-Smirnov tests (SOKAL, ROHLF 1995) were carried out to assess the statistical significance of differences between these groups.

Phylogenetic analysis·

Phylogenetic trees were constructed for the three groups of orthologous sequences extracted from the sets of 1521 orthologs present in all the 7 genomes of γ-Proteobacteria:

– a group of 191 orthologous sequences lying in all the analyzed genomes on the leading strand;

– a group of 38 orthologous sequences lying in all the analyzed genomes on the lagging strand;

– a group of 8 orthologous sequences, which in 6 genomes lie on the leading strand but in *E. coli* EDL933 on the lagging strand (2 genes of *E. coli* EDL933 and their orthologs that meet this criterion and were subjected to horizontal transfer according to Horizontal Gene Transfer Database (HGT-DB) (GARCIA-VALLVE et al. 2003) were excluded from analysis. These orthologs are classified in the follow-

ing COGs: COG0217 (uncharacterized ACR), COG0236 (acyl carrier protein), COG0322 (nuclease subunit of the excinuclease complex), COG0494 (NTP pyrophosphohydrolases including oxidative damage repair enzymes), COG0740 (protease subunit of ATP-dependent Clp proteases), COG1028 (dehydrogenases with different specificities related to short-chain alcohol dehydrogenases), COG1130 (ABC-type sugar/spermidine/putrescine/iron/thiamine transport systems, ATPase component).

In the construction of phylogenetic trees, the Dayhoff (DAYHOFF et al. 1978) and JTT (JONES et al. 1992) models of amino acid substitutions were used, as implemented in the TREE-PUZZLE program (SCHMIDT et al. 2002).

Evolutionary distances between 16S rRNA sequences (measured by the number of substitutions per site) were calculated by the MEGA 2.1 program (KUMAR et al. 1993), assuming the Tamura-Nei model of nucleotide substitutions (TAMURA, NEI 1993). The 16S rRNA tree was built by the neighbor-joining, minimum evolution and maximum parsimony methods with the MEGA 2.1 program.

Results and discussion

The first stage of our study was performed with pairs of closely-related genomes in the COG database (three levels of relations: intraspecific, interspecific and intergeneric). Orthologs found in each pair of genomes were divided into three groups: located on leading strands, on lagging strands, and on different strands – trans-orthologs. For each pair we measured the distance as the mean number of amino acid substitutions per site between the closest orthologs (unidirectional best hits). For almost all analyzed pairs of the closest genomes, the highest divergence was observed for trans-orthologs, which suggests that translocation of a sequence between differently replicating DNA strands accelerates the accumulation of mutations inside it (Table 1). Moreover, we have found that the divergence of sequences located on lagging strands is usually higher than the divergence of sequences located on leading strands. We obtained similar results by drawing pairwise comparisons for seven genomes belonging to the γ-Proteobacteria group. In Figures 1A, B, C (data for 7 γ-Proteobacteria genomes) and Figures 2A, B, C (data for comparisons of *E. coli* EDL933 with 13 genomes), the divergence values of the three classes of orthologs were plotted against the phylogenetic distances between the compared genomes, counted on the basis of the divergence of 16S rRNA genes. The divergence of the orthologs lying in both genomes on the same DNA strand is linearly correlated with the distance measured by divergence of 16S rRNA genes (correlation coefficient is statistically significant for both groups), but note that the value $b > 0$ in the linear regression equation $y = ax+b$ suggests that in fact there is a deviation from linear relation between the rate of accumulation of mutation and time elapsed after

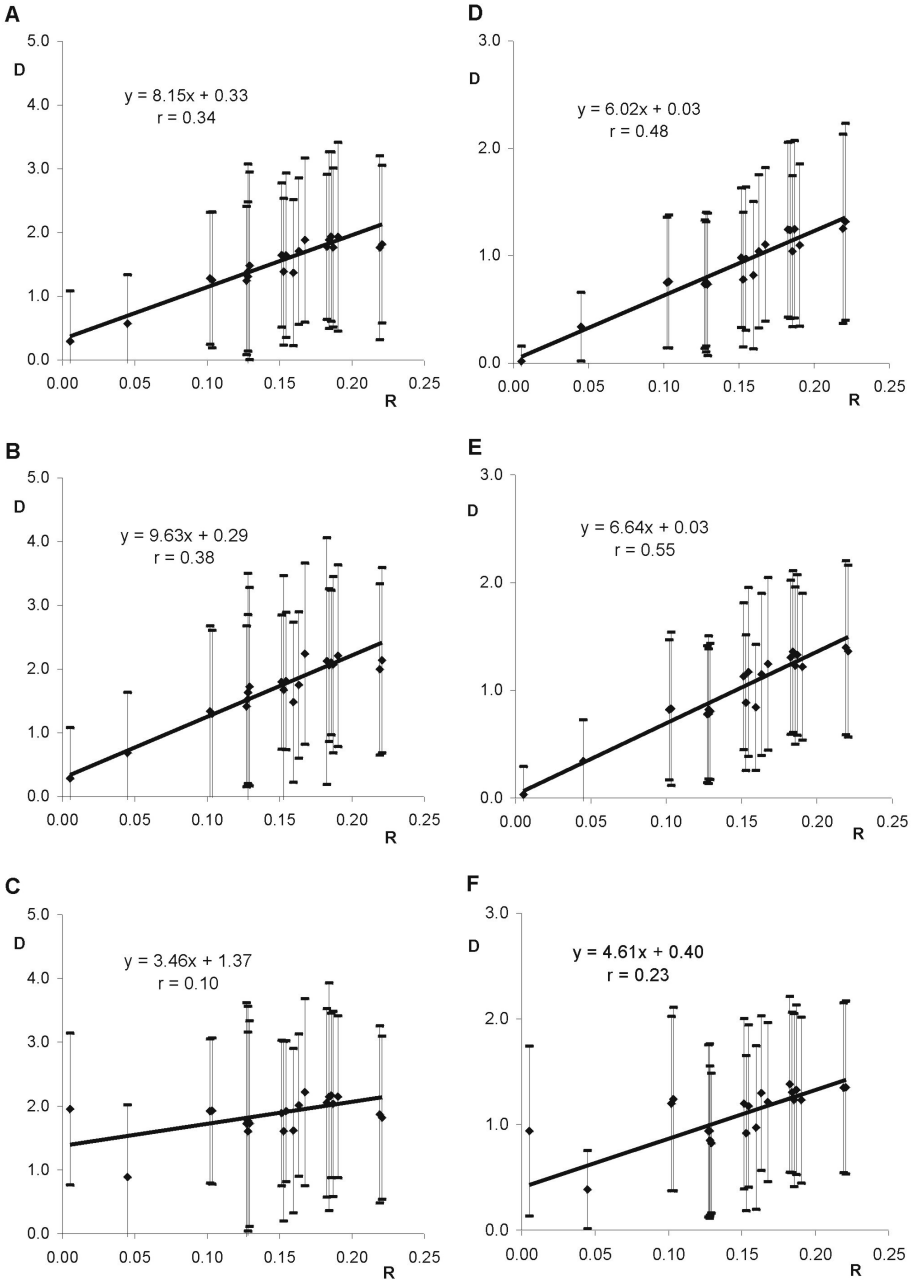


Figure 1. Relation between the divergence (D) of orthologs and the phylogenetic distance (R) measured by 16S rRNA performed for three groups of orthologs: lying on the leading strand (top row), lying on the lagging strand (middle row) and trans-orthologs (bottom row). Results for orthologs chosen by the unidirectional best hit rule are in the left column and for orthologs chosen by the bidirectional best hit rule in the right column. Data obtained from pairwise comparison of 7 genomes belonging to γ-Proteobacteria. Bars represent standard deviation.

Table 1. The mean evolutionary distances for three groups of orthologs (chosen by unidirectional best hit rule) and statistical significance of differences between them.

Pair of compared genomes	Mean distance between orthologs and *standard deviation* (number of orthologs for given pair of genomes in parentheses)			Statistical significance of differences between distances:		
	on leading strand (1)	on lagging strand (2)	switched DNA strand (3)	(1)-(2)	(1)-(3)	(2)-(3)
1	2	3	4	5	6	7
Bb – Tp	(310) **1.529** ± *0.860*	(78) **1.776** ± *0.822*	(193) **1.916** ± *1.145*	**	**	n
Bh – Bs	(2224) **1.136** ± *0.916*	(310) **1.369** ± *0.939*	(875) **1.640** ± *0.965*	**	**	**
Cc – Mlt	(1544) **1.705** ± *1.114*	(1030) **1.911** ± *1.232*	(2130) **1.901** ± *1.206*	**	**	n
Cp – Ct	(339) **0.468** ± *0.271*	(270) **0.542** ± *0.316*	(37) **1.079** ± *1.448*	**	**	*
EcE – EcK	(1843) **0.295** ± *0.792*	(1423) **0.284** ± *0.799*	(391) **1.953** ± *1.190*	*	**	**
EcE – Hi	(811) **1.480** ± *1.471*	(559) **1.723** ± *1.557*	(1148) **1.725** ± *1.613*	**	**	n
EcE – Pa	(1777) **1.648** ± *1.130*	(1136) **1.797** ± *1.051*	(2320) **1.888** ± *1.140*	**	**	**
EcE – Pm	(968) **1.312** ± *1.171*	(603) **1.529** ± *1.326*	(1166) **1.768** ± *1.798*	**	**	**
EcE – Vc	(1519) **1.284** ± *1.036*	(900) **1.337** ± *1.342*	(1350) **1.922** ± *1.133*	n	**	**
EcE – Xf	(901) **1.884** ± *1.384*	(450) **2.061** ± *1.196*	(1075) **2.145** ± *1.785*	**	**	n
EcK – Hi	(700) **1.385** ± *1.690*	(496) **1.631** ± *1.871*	(954) **1.601** ± *1.561*	**	**	n
EcK – Pa	(1576) **1.646** ± *1.287*	(1054) **1.812** ± *1.077*	(2086) **1.919** ± *1.104*	**	**	**
EcK – Pm	(834) **1.248** ± *1.165*	(524) **1.414** ± *1.262*	(1024) **1.723** ± *1.900*	**	**	**
EcK – Vc	(1336) **1.259** ± *1.064*	(815) **1.298** ± *1.313*	(1162) **1.923** ± *1.148*	n	**	**

1	2	3	4	5	6	7
EcK – Xf	(762) **1.767** ± *1.246*	(394) **2.069** ± *1.386*	(902) **2.033** ± *1.451*	**	**	n
Hi – Pa	(946) **1.939** ± *1.329*	(590) **2.103** ± *1.135*	(1381) **2.165** ± *1.291*	**	**	n
Hi – Pm	(586) **0.572** ± *0.766*	(410) **0.685** ± *0.951*	(646) **0.887** ± *1.134*	n	**	**
Hi – Vc	(693) **1.388** ± *1.149*	(425) **1.675** ± *1.794*	(952) **1.607** ± *1.409*	**	**	n
Hi – Xf	(462) **1.764** ± *1.441*	(229) **1.996** ± *1.343*	(571) **1.869** ± *1.386*	**	*	n
Hp – HpJ	(603) **0.132** ± *0.523*	(424) **0.135** ± *0.503*	(101) **0.638** ± *1.025*	n	**	**
Ll – Sp	(1012) **1.088** ± *0.907*	(152) **1.102** ± *0.778*	(351) **1.731** ± *0.978*	n	**	**
Ml – Mt	(1014) **0.763** ± *1.074*	(474) **0.960** ± *1.137*	(523) **1.717** ± *1.208*	**	**	**
NmM – NmZ	(796) **0.062** ± *0.302*	(680) **0.075** ± *0.361*	(88) **1.020** ± *1.277*	**	**	**
Pa – Pm	(602) **0.337** ± *0.546*	(513) **0.444** ± *0.788*	(338) **0.882** ± *1.064*	**	**	n
Pa – Vc	(1130) **1.937** ± *1.478*	(643) **2.210** ± *1.426*	(1542) **2.146** ± *1.270*	n	**	**
Pa – Xf	(1716) **1.712** ± *1.148*	(959) **1.750** ± *1.148*	(1948) **2.013** ± *1.114*	**	**	n
Pab – Ph	(1182) **1.884** ± *1.288*	(562) **2.241** ± *1.421*	(1481) **2.220** ± *1.467*	n	**	**
Pm – Vc	(850) **1.372** ± *1.146*	(506) **1.481** ± *1.258*	(997) **1.614** ± *1.288*	n	**	**
Pm – Xf	(509) **1.820** ± *1.234*	(233) **2.139** ± *1.455*	(659) **1.818** ± *1.278*	**	n	**
Vc – Xf	(824) **1.778** ± *1.139*	(358) **2.126** ± *1.935*	(930) **2.053** ± *1.478*	**	**	n

The distances between orthologs are expressed by the mean number of amino acid substitutions per site between the two genomes. Statistical significance of differences between the distances was analyzed by Mann-Whitney U and Kolmogorov-Smirnov tests. The significance level for the differences is: ** ($p < 0.01$), * ($0.01 \; p < 0.05$), n (not significant, $p \geq 0.05$). For genome name abbreviations, see Material and methods.

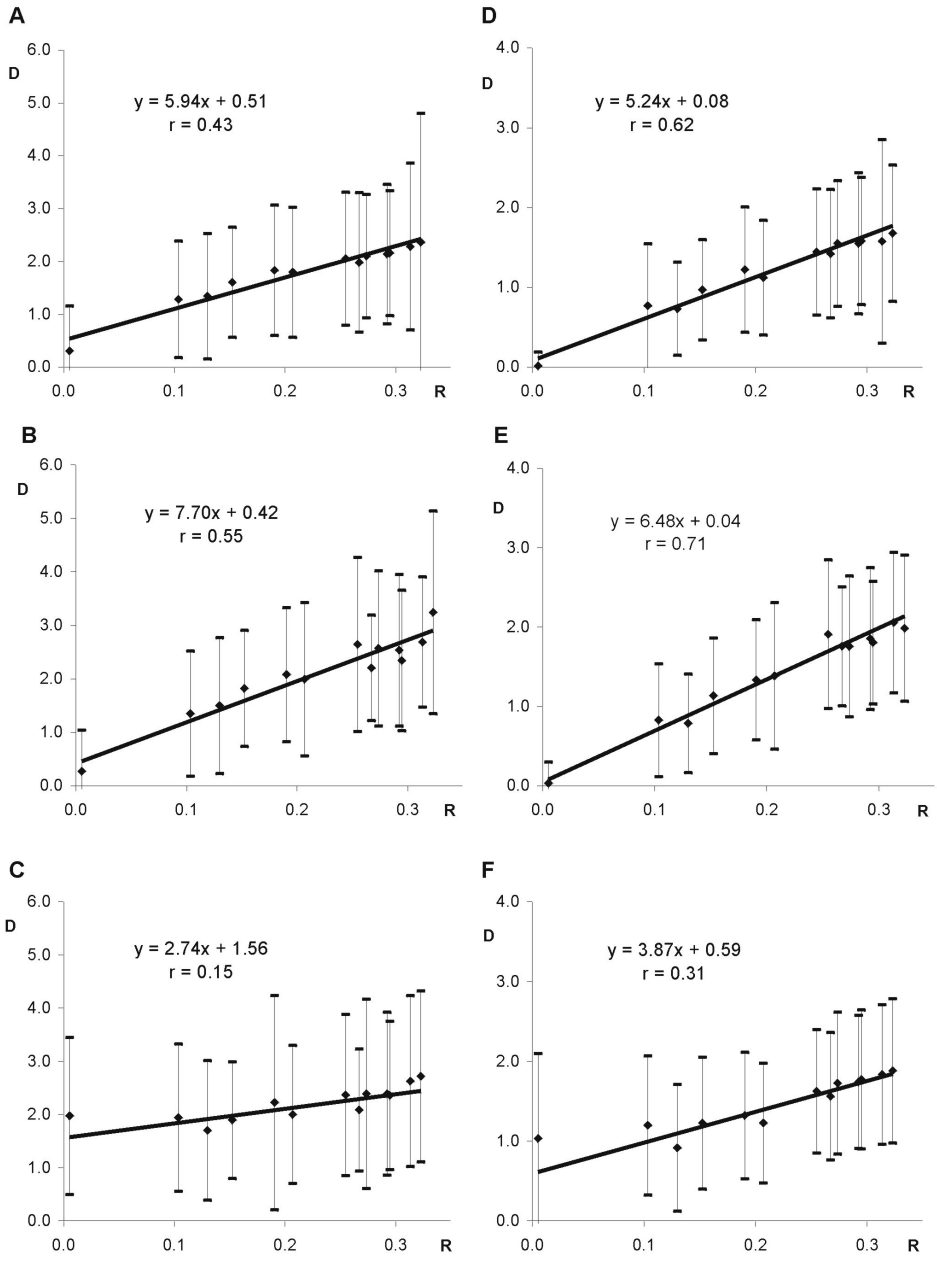


Figure 2. Relation between the divergence (D) of orthologs and the phylogenetic distance (R) measured by 16S rRNA performed for three groups of orthologs: lying on the leading strand (top row), lying on the lagging strand (middle row) and trans-orthologs (bottom row). Results for orthologs chosen by the unidirectional best hit rule are in the left column and for orthologs chosen by the bidirectional best hit rule in the right column. Data obtained from comparisons of *E. coli* EDL933 with 13 genomes. Bars represent standard deviation.

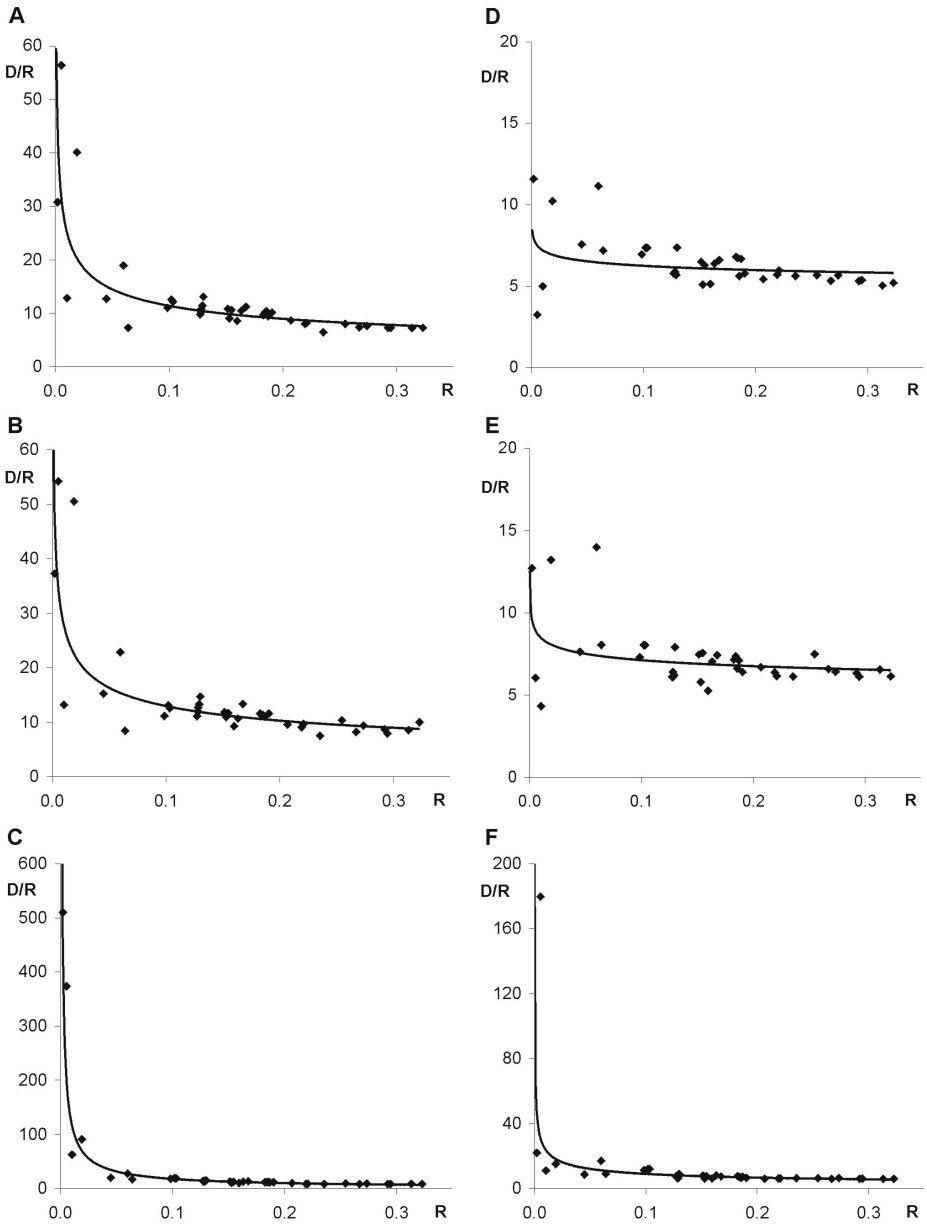


Figure 3. Relation between the divergence rate (D/R) of orthologs and the phylogenetic distance (R) measured by 16S rRNA performed for three classes of orthologs: lying on the leading strand (top row), lying on the lagging strand (middle row) and trans-orthologs (bottom row). Results for orthologs chosen by the unidirectional best hit rule are in the left column and for orthologs chosen by the bidirectional best hit rule in the right column. Data obtained from pairwise comparison of 7 genomes of γ-Proteobacteria, comparison of pairs of closely-related genomes in the COG database, and comparison of *E. coli* EDL933 with 13 genomes.

Table 2. The mean evolutionary distances for three groups of orthologs (chosen by bidirectional best hit rule) and statistical significance of differences between them

Pair of compared genomes	Mean distance between orthologs and *standard deviation* (number of orthologs for given pair of genomes in parentheses)			Statistical significance of differences between distances:		
	on leading strand (1)	on lagging strand (2)	switched DNA strand (3)	(1)-(2)	(1)-(3)	(2)-(3)
1	2	3	4	5	6	7
Bb-Tp	(256) **1.324** ± *0.647*	(55) **1.444** ± *0.559*	(136) **1.488** ± *0.697*	*	*	n
Bh-Bs	(1401) **0.667** ± *0.462*	(171) **0.837** ± *0.501*	(301) **1.016** ± *0.588*	**	**	**
Cc-Mlt	(623) **0.959** ± *0.553*	(368) **1.029** ± *0.583*	(724) **1.144** ± *0.648*	n	**	**
Cp-Ct	(335) **0.460** ± *0.259*	(264) **0.516** ± *0.266*	(31) **0.578** ± *0.277*	**	*	n
EcE-EcK	(1569) **0.017** ± *0.140*	(1234) **0.032** ± *0.262*	(16) **0.939** ± *0.804*	**	**	**
EcE-Hi	(484) **0.734** ± *0.664*	(278) **0.805** ± *0.633*	(545) **0.824** ± *0.661*	**	**	n
EcE-Pa	(851) **0.983** ± *0.649*	(480) **1.131** ± *0.681*	(756) **1.198** ± *0.806*	**	**	n
EcE-Pm	(619) **0.739** ± *0.579*	(340) **0.783** ± *0.605*	(569) **0.938** ± *0.824*	*	**	*
EcE-Vc	(933) **0.750** ± *0.610*	(583) **0.821** ± *0.650*	(460) **1.197** ± *0.826*	*	**	**
EcE-Xf	(494) **1.239** ± *0.823*	(208) **1.359** ± *0.751*	(472) **1.304** ± *0.759*	**	**	n
EcK-Hi	(462) **0.754** ± *0.649*	(272) **0.822** ± *0.685*	(508) **0.848** ± *0.707*	**	**	n
EcK-Pa	(761) **0.973** ± *0.667*	(437) **1.171** ± *0.784*	(661) **1.175** ± *0.769*	**	**	n
EcK-Pm	(564) **0.735** ± *0.599*	(301) **0.779** ± *0.636*	(536) **0.939** ± *0.814*	n	**	**
EcK-Vc	(859) **0.760** ± *0.620*	(538) **0.830** ± *0.712*	(415) **1.241** ± *0.868*	**	**	**

1	2	3	4	5	6	7
EcK-Xf	(462) **1.248** ± *0.826*	(188) **1.329** ± *0.745*	(450) **1.329** ± *0.802*	*	**	n
Hi-Pa	(426) **1.043** ± *0.705*	(207) **1.230** ± *0.729*	(502) **1.232** ± *0.819*	**	**	n
Hi-Pm	(504) **0.340** ± *0.320*	(332) **0.343** ± *0.385*	(488) **0.384** ± *0.370*	n	**	n
Hi-Vc	(438) **0.779** ± *0.627*	(233) **0.886** ± *0.630*	(556) **0.918** ± *0.734*	**	**	n
Hi-Xf	(314) **1.252** ± *0.881*	(148) **1.397** ± *0.807*	(380) **1.348** ± *0.803*	**	*	n
Hp-HpJ	(578) **0.051** ± *0.147*	(402) **0.044** ± *0.071*	(72) **0.113** ± *0.367*	n	*	*
Ll-Sp	(711) **0.685** ± *0.466*	(96) **0.720** ± *0.419*	(127) **1.100** ± *0.741*	*	**	**
Ml-Mt	(737) **0.194** ± *0.158*	(295) **0.251** ± *0.330*	(131) **0.284** ± *0.258*	**	**	*
NmM-NmZ	(747) **0.023** ± *0.101*	(626) **0.025** ± *0.115*	(25) **0.044** ± *0.096*	**	**	**
Pab-Ph	(552) **0.210** ± *0.222*	(459) **0.217** ± *0.217*	(221) **0.289** ± *0.444*	n	n	n
Pa-Pm	(508) **1.099** ± *0.757*	(252) **1.220** ± *0.680*	(559) **1.231** ± *0.786*	**	**	n
Pa-Vc	(852) **1.043** ± *0.714*	(450) **1.149** ± *0.752*	(596) **1.297** ± *0.733*	**	**	**
Pa-Xf	(541) **1.106** ± *0.715*	(212) **1.246** ± *0.802*	(497) **1.212** ± *0.753*	**	**	n
Pm-Vc	(548) **0.819** ± *0.686*	(308) **0.843** ± *0.585*	(567) **0.971** ± *0.775*	n	**	*
Pm-Xf	(337) **1.317** ± *0.917*	(144) **1.365** ± *0.798*	(444) **1.351** ± *0.818*	n	n	n
Vc-Xf	(472) **1.243** ± *0.813*	(183) **1.307** ± *0.716*	(452) **1.381** ± *0.833*	n	**	n

The distances between orthologs are expressed by the mean number of amino acid substitutions per site between the two genomes. Statistical significance of differences between the distances was analyzed by Mann-Whitney U and Kolmogorov-Smirnov tests. The significance level for the differences is: ** ($p < 0.01$), * ($0.01 \; p < 0.05$), n (not significant, $p \geq 0.05$). For genome name abbreviations, see Material and methods.

Table 3. The mean evolutionary distances for three groups of orthologs (chosen by bidirectional best hit rule and occupying the conserved positions in two genomes) and statistical significance of differences between them.

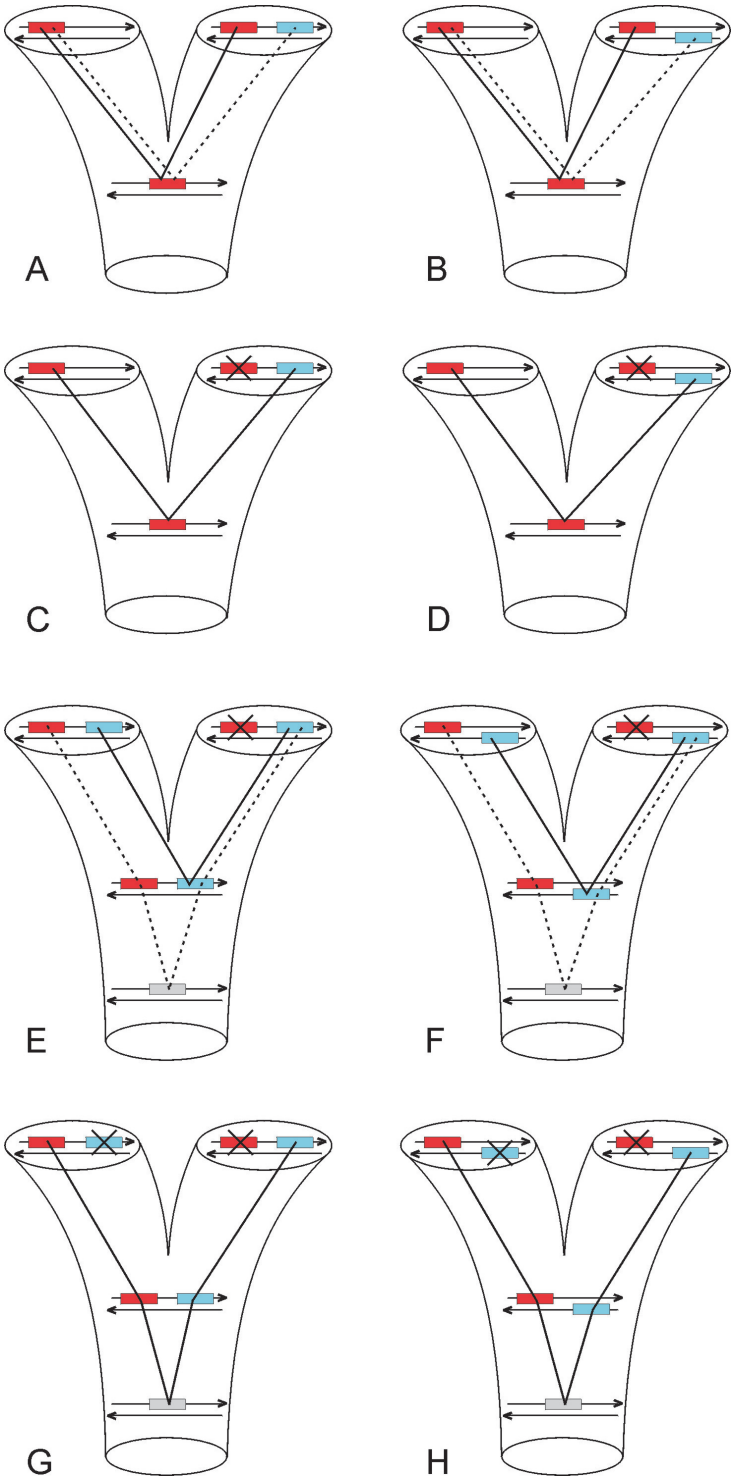
Pair of compared genomes	Mean distance between orthologs and *standard deviation* (number of orthologs for given pair of genomes in parentheses)			Statistical significance of differences between distances:		
	on leading strand (1)	on lagging strand (2)	switched DNA strand (3)	(1)-(2)	(1)-(3)	(2)-(3)
Bh-Bs	(976) **0.558** ± *0.343*	(101) **0.642** ± *0.297*	(30) **0.855** ± *0.574*	**	**	
Cp-Ct	(331) **0.458** ± *0.258*	(259) **0.510** ± *0.262*	(28) **0.585** ± *0.283*	**	*	n
EcE-EcK	(1561) **0.013** ± *0.097*	(1226) **0.025** ± *0.227*	-	**	-	-
NmM-NmZ	(742) **0.020** ± *0.055*	(620) **0.025** ± *0.116*	(21) **0.026** ± *0.015*	**	**	**

The distances between orthologs are expressed by the mean number of amino acid substitutions per site between the two genomes. Statistical significance of differences between the distances was analyzed by Mann-Whitney U and Kolmogorov-Smirnov tests. The significance level for the differences is: ** ($p < 0.01$), * ($0.01 \; p < 0.05$), n (not significant, $p \geq 0.05$). For genome name abbreviations, see Material and methods.

the divergence of species seen for pairs of the closest genomes. Such an effect is much more strongly pronounced for pairs of sequences that switched the DNA strand. This is a paradox, because the divergence rate for the closest pairs is higher than for the pairs of genomes at larger distances. This is conspicuous in the plot where the divergence rate (i.e. divergence of the analyzed sequences divided by the distance estimated on the basis of analysis of 16S rRNA) is plotted against the phylogenetic distance (Figures 3A, B, C – data for all three sets of genomes). For the closest genomes the divergence rate is several times higher than for the most distant genomes and it is most evident for the sequences that switched the DNA strand. The same analysis performed with the set of the homologs that are bidirectional best hits is shown in Table 2, Figures 1D, E, F, Figures 2D, E, F and Figures 3D, E, F. Differences in divergence between the three classes of orthologs according to their location on DNA strands are still observed, although the absolute values of divergence are smaller than in the first approach. Moreover, the shift of the linear regression lines on the y-axis has disappeared for the classes of orthologs that have not switched the DNA strand (value $b \approx 0$ in the regression line equation, Figures 1D, E and Figures 2D, E). However, the shift is still observed for the set of trans-orthologs (Figure 1F and Figure 2F) and the divergence rate in this group is the highest for the closely related genomes (Figure 3F).

To eliminate the possibility of affecting our results by such paralogs and horizontally transferred genes, we have performed our studies only with homologs in the conserved positions. That has not changed the previous results (Table 3). If we put the genes staying in both genomes on the same strand (leading or lagging) into one set, there is a statistically significant difference (for *Chlamydia* genomes $p = 0.05$ and for *Bacillus* and *Neisseria* genomes $p < 0.005$) in the divergence of these genes and the genes that switched the DNA strands.

This result suggests that there is a strong bias in representation of highly diverged sequences lying on differently replicating strands, and a relatively larger fraction of such sequences is found in the closely related genomes. To consider possible explanations of this finding, we have drawn in Figure 4 the possible pathways of gene evolution, including gene duplications and inversions in the two diverging genomes (note that by "inversion" we understand the translocation from the leading to the lagging strand or *vice versa*). The case of evolution of orthologs staying on the same DNA strand during evolution is shown in the left panel of Figure 4. In the right panel, we presented the possible situations when some of the orthologs switched the DNA strand. Lines indicate the history of genes and distances measured between orthologs in phylogenetic analysis. Solid lines show comparisons that are found in the bidirectional best hits approach. Dashed lines indicate comparisons found only in the unidirectional best hits method, which disappear when the bidirectional best hits rule is applied. Figures 4A and B illustrate two possible duplications of a gene in one of the genomes: on the same strand (A) and duplication with inversion on the other strand (B).


A
B
C
D
E
F
G
H

If the bidirectional best hits rule is used, the pair of trans-orthologs disappears from the comparison because of a higher divergence of the inverted copy. If the maternal copy of a gene decays in one of the genomes (Figures 4C, D), the trans-orthologs are found in the analysis when the bidirectional best hits rule is applied. This case may be also considered a simple translocation of a gene on the chromosome with and without inversion. Figures 4E and F show the case when duplications of genes had occurred before the divergence of the taxa, but one copy of the gene vanished in one of the genomes. Unidirectional distances between such orthologs could show false distances between genomes, because they correspond with the time of duplication and separation of gene copies rather than to taxa. These misleading comparisons disappear when the bidirectional best hits rule is applied and only distances between nearest orthologs are measured. However, if different paralogs disappear in the two lineages (Figures 4G, H), these comparisons may be detected in the phylogenetic analysis. The above arguments may explain the nonzero value of *b* of the linear regression lines for the unidirectional orthologs analysis for all classes of orthologs. When the bidirectional best hits rule is applied, the shift is still observed for trans-orthologs and the divergence rate is still the highest for the closely related genomes. It is clear that for the closely related genomes, most of orthologs have not changed their locations on chromosomes. Small fractions of translocated sequences (without inversions) do not change significantly the results of analysis. This is not the case for trans-orthologs. The fraction of these sequences among translocated sequences is always significant. Furthermore, the fraction of sequences whose history is shown in Figure 4H could be also significant. It is easy to predict that in this class of sequences, the transposons can be found. In fact, in case of intraspecific comparisons of *E. coli* strains, 8 out of 16 found trans-orthologs belong to transposases or insertion elements. This effect becomes "diluted" as the distance between compared taxa grows. Nevertheless, there is also a possibility that the cases described in Figure 4D, when a higher divergence of sequences after inversion is observed, could affect the results of comparisons. The sequence duplicated with inversion is subject to a higher mutational pressure than the maternal copy staying on the same

Figure 4. The possible pathways of gene evolution, including gene duplications, inversions and disappearance in two diverging genomes. The cases of evolution of orthologs staying on the same DNA strand during evolution are shown in the left panel. The cases of evolution including duplication/translocation of a gene on the other DNA strand (inversion) are presented in the right panel. Lines indicate the history of genes and distances measured between orthologs in the phylogenetic analysis. Solid lines show comparisons between genes that are found in the bidirectional best hits approach. Dashed lines indicate comparisons found only in the unidirectional best hits method, which disappear when the bidirectional best hits rule is applied. Two gene copies are indicated by red and blue boxes, their ancestor is in gray. Antiparallel arrows mean two differently replicating strands. Gene disappearance is indicated by X.

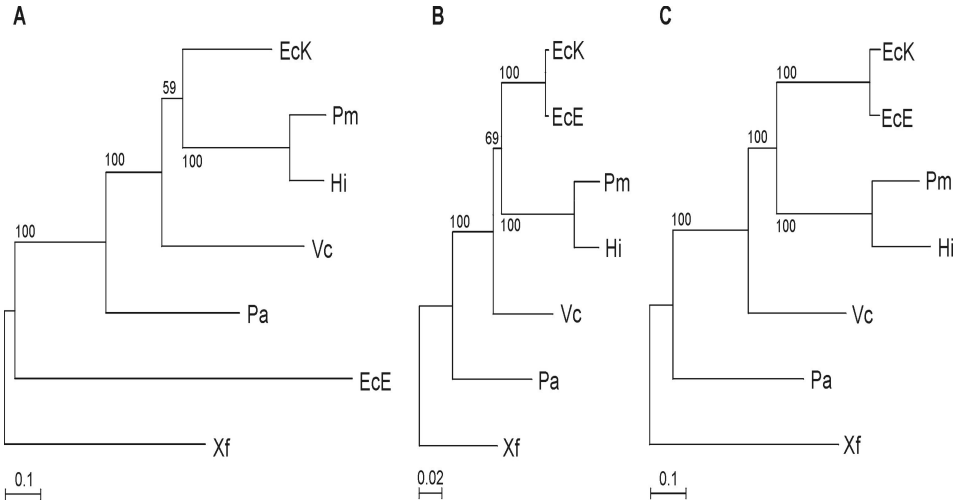


Figure 5. Phylogenetic trees of 7 taxa belonging to γ-Proteobacteria constructed for: (A) group of 8 orthologous sequences, which in 6 genomes lies on the leading strand but in *E. coli* EDL933 (EcE) on the lagging strand; (B) nucleotide sequences of 16S rRNA; (C) group of 191 orthologous sequences lying in all analyzed genomes on the leading strand. Trees on (A) and (C) were constructed by TREE-PUZZLE 5.0 program assuming the JTT model of amino acid substitutions. 16S rRNA tree was built by the neighbor-joining method with the MEGA 2.1 program assuming Tamura-Nei model of nucleotide substitutions. Bootstrap values (at nodes) were calculated by the analysis of 1000 replicates. The scale bar for (A) and (C) represents the number of amino acid substitutions per site and for (B) the number of nucleotide substitutions per site. Genome name abbreviations: EcK = *Escherichia coli* K12-MG1655; EcE = *E. coli* O157:H7 EDL933; Hi = *Haemophilus influenzae*; Pm = *Pasteurella multocida*; Pa = *Pseudomonas aeruginosa*; Vc = *Vibrio cholerae*; Xf = *Xylella fastidiosa*.

strand, because its nucleotide composition is more distant from the equilibrium with the new mutational pressure than that of the sequence staying on the same strand. If the function of the translocated sequence is not indispensable for the survival of the organism, it is released from the strong selection pressure. This sequence may evolve into a gene fulfilling another function or may become a pseudogene because of accumulation of too many substitutions. As these sequences diverge very fast, they eventually disappear from the comparative studies of more distant genomes. As a result, only sequences accepted by selection persist for long evolutionary distances and they dominate in the set of more distant genomes. Thus, relative differences between divergence of orthologs lying on the same strand and trans-orthologs decrease.

If the last effect – generation of pseudogenes – is true, sequences shortened by mutations should be observed especially in the class of trans-orthologs. Actually, we found for 11 pairs of closely-related genomes, that the average length of alignments of trans-orthologs equals 312.0, while the average length of alignments

of orthologs that have not changed DNA strand is 324.7, and the difference is statistically significant ($p < 0.0001$, Kolmogorov-Smirnov test). These data are found by the bidirectional analysis and the difference is much more profound for the unidirectional analysis. For some more distant genomes this difference is still observed but it is smaller and not statistically significant (data not shown). One could argue that the length of alignments is negatively correlated with divergence, which could explain the observed difference in the length of orthologs. However, the correlation between the length of alignments and divergence is very low: –0.07 (this is the average correlation coefficient counted from values obtained for each pair of compared genomes). OCHMAN (2002) found that a substantial fraction of hypothetical open reading frames are actually short. This suggests that many of them are not functional genes. Furthermore, MIRA et al. (2001), who analyzed known pseudogenes in a broad taxonomic range of bacteria, observed that in every case deletions are more frequent than insertions, which results in shortening of inactive genes. A substantial fraction of short ORFs (assumed non-coding ORFs) have been observed in the genome of the yeast *Saccharomyces cerevisiae* (ANDRADE et al. 1997, DAS et al. 1997, MACKIEWICZ et al. 2002).

We do not exclude that the horizontally transferred genes may be found in the set of orthologs lying on different strands. On the other hand, it is unlikely that all these orthologs were acquired by HGT. Assuming that HGT strongly influences our results, we should accept that horizontally transferred genes are preferably found in closely related genomes and they are preferentially found in the set of orthologs that switched DNA strands, which in our opinion is rather unlikely. However, there should be the same probability that a gene may be transferred on the same and on the differently replicating strand, so the contribution of transferred genes to sets of orthologs lying on the same strand and on the different replicating strand should be the same. As we are interested in relative comparison of divergence of different sets but not absolute values, HGT has little effect on our results. Moreover, it does not preclude that inverted genes accumulate more substitutions irrespective of their origin.

The obtained results are important for some phylogenetic studies. If for such studies the sequences of trans-orthologs are chosen (accidentally or not), the obtained phylogenetic distances between taxa could be false or the phylogenetic tree could show a false history of divergence. To exemplify this effect, we have constructed a phylogenetic tree based on 8 orthologous sequences present in all the 7 genomes belonging to γ-Proteobacteria (Figure 5A). In 6 genomes, all these orthologs lie on the leading strand, but in *E. coli* EDL933 (EcE) they are located on the lagging strand, which indicates that in this genome these orthologs switched their strand. The obtained tree has a different topology than 16S rRNA trees constructed by three different methods giving the same topology (Figure 5B – only the neighbor-joining tree is shown). Furthermore, the branch of the tree leading to *E. coli* EDL933 is longer, indicating a higher divergence rate in this taxon. On the other hand, the trees obtained for the orthologs lying in all the ana-

lyzed genomes on the same strand (leading or lagging) have the same topology as the 16S rRNA tree (Figure 5C the lagging strand orthologs tree is not shown). There were no differences in topology of respective trees for orthologous sequences when the Dayhoff and JTT models of amino acid substitutions were used. The influence of fast-evolving species on topology of phylogenetic trees was reviewed by PHILIPPE and LAURENT (1998).

Conclusions

Translocation of a coding sequence to a differently replicating DNA strand often causes an immediate mutagenic effect on the sequence and acceleration of its divergence. The sequences translocated between the DNA strands showing a high divergence rate may belong to: (1) paralogs or other fast-evolving genes under weak selection; or (2) pseudogenes that will probably be eliminated from the genome during further evolution; or (3) genes whose history after divergence is longer than the history of the genomes in which they are found. Some of these sequences may survive and if they are accepted by selection, they may gain new functions and enable adaptation of microorganisms to the changing environment. This may accelerate the evolution of bacteria. In conclusion, it is important to use in phylogenetic analyses the sequences evolving at similar rates if one wants to get comparable distances for the whole phylogenetic tree.

Acknowledgments. The work was supported by the grants: 1016/S/IMi/03, KBN 3PO4A 004-22. Additionally, M.K. was supported by the Foundation for Polish Science.

REFERENCES

ANDERSSON J.O., ANDERSSON S.G. (1999). Genome degradation is an ongoing process in *Rickettsia*. Mol. Biol. Evol. 16: 1178-1191.

ANDERSSON J.O., ANDERSSON S.G. (2001). Pseudogenes, junk DNA, and the dynamics of *Rickettsia* genomes. Mol. Biol. Evol. 18: 829-839.

ANDERSSON S.G., ZOMORODIPOUR A., ANDERSSON J.O., SICHERITZ-PONTEN T., ALSMARK U.C., PODOWSKI R.M., NASLUND A.K., ERIKSSON A.S., WINKLER H.H., KURLAND C.G. (1998). The genome sequence of *Rickettsia prowazekii* and the origin of mitochondria. Nature 396: 133-140.

ANDRADE M.A., DARUVAR A., CASARI G., SCHNEIDER R., TERMIER M., SANDER C. (1997). Characterization of new proteins found by analysis of short open reading frames from the full yeast genome. Yeast 13: 1363-1374.

BLATTNER F.R., PLUNKETT G. 3rd, BLOCH C.A., PERNA N.T., BURLAND V., RILEY M., COLLADO-VIDES J., GLASNER J.D., RODE C.K., MAYHEW G.F. et al. (1997). The complete genome sequence of *Escherichia coli* K-12. Science 277: 1453-1462.

CEBRAT S., STAUFFER D. (2002). Monte Carlo simulation of genome viability. J. Appl. Genet. 43: 391-395.

COLE S.T., EIGLMEIER K., PARKHILL J., JAMES K.D., THOMSON N.R., WHEELER P.R., HONORE N., GARNIER T., CHURCHER C., HARRIS D. et al. (2001). Massive gene decay in the leprosy bacillus. Nature 409: 1007-1011.

DAS S., YU L., GAITATZES C., ROGERS R., FREEMAN J., BIENKOWSKA J., ADAMS R.M., SMITH T.F., LINDELIEN J. (1997). Biology's new Rosetta stone. Nature 385: 29-30.

DAYHOFF M.O., SCHWARTZ R.M., ORCUTT B.C. (1978). A model of evolutionary change in proteins. In: Atlas of protein sequence and structure. (M.O. Dayhoff, ed.) Natl. Biomed. Res. Found., Washington, DC, 5 (Suppl. 3): 345-352.

DELORME C., GODON J.J., EHRLICH S.D., RENAULT P. (1993). Gene inactivation in *Lactococcus lactis*: histidine biosynthesis. J. Bacteriol. 175: 4391-4399.

FELSENSTEIN J. (1993). PHYLIP: Phylogeny Inference Package, version 3.5c. Distributed by the author. Department of Genetics, University of Washington, Seattle.

FITCH W.M. (1970). Distinguishing homologous from analogous proteins. Syst. Zool. 19: 99-113.

FRANCINO M.P., OCHMAN H. (1997). Strand asymmetries in DNA evolution. Trends Genet. 13: 240-245.

FRANK A.C., LOBRY J.R. (1999). Asymmetric substitution patterns: a review of possible underlying mutational or selective mechanisms. Gene 238: 65-77.

FRASER C.M., CASJENS S., HUANG W.M., SUTTON G.G., CLAYTON R., LATHIGRA R., WHITE O., KETCHUM K.A., DODSON R., HICKEY E.K. et al. (1997). Genomic sequence of a Lyme disease spirochaete, *Borrelia burgdorferi*. Nature 390: 580-586.

FRASER C.M., NORRIS S.J., WEINSTOCK G.M., WHITE O., SUTTON G.G., DODSON R., GWINN M., HICKEY E.K., CLAYTON R., KETCHUM K.A. et al. (1998). Complete genome sequence of *Treponema pallidum*, the syphilis spirochete. Science 281: 375-388.

FREEMAN J.M., PLASTERER T.N., SMITH T.F., MOHR S.C. (1998). Patterns of genome organization in bacteria. Science 279: 1827.

GARCIA-VALLVE S., GUZMAN E., MONTERO M.A., ROMEU A. (2003). HGT-DB: a database of putative horizontally transferred genes in prokaryotic complete genomes. Nucleic Acids Res. 31: 187-189.

GOJOBORI T., LI W.-H., GRAUR D. (1982). Patterns of nucleotide substitution in pseudogenes and functional genes. J. Mol. Evol. 18: 360-369.

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

HARRISON P.M., GERSTEIN M. (2002). Studying genomes through the aeons: protein families, pseudogenes and proteome evolution. J. Mol. Biol. 318: 1155-1174.

HOMMA K., FUKUCHI S., KAWABATA T., OTA M., NISHIKAWA K. (2002). A systematic investigation identifies a significant number of probable pseudogenes in the *Escherichia coli* genome. Gene 294: 25-33.

HUYNEN M.A., VAN NIMWEGEN E. (1998). The frequency distribution of gene family sizes in complete genomes. Mol. Biol. Evol. 15: 583-589.

JONES D.T., TAYLOR W.R., THORNTON J.M. (1992). The rapid generation of mutation data matrices from protein sequences. Comput. Appl. Biosci. 8: 275-282.

KONDRASHOV F.A., ROGOZIN I.B., WOLF Y.I., KOONIN E.V. (2002). Selection in the evolution of gene duplications. Genome Biol. 3(2): research0008.

KOONIN E.V., TATUSOV R.L., GALPERIN M.Y. (1998). Beyond complete genomes: from sequence to structure and function. Curr. Opin. Struct. Biol. 8: 355-363.

KOWALCZUK M., MACKIEWICZ P., MACKIEWICZ D., NOWICKA A., DUDKIEWICZ M., DUDEK M.R., CEBRAT S. (2001a). DNA asymmetry and the replicational mutational pressure. J. Appl. Genet. 42: 553-577.

KOWALCZUK M., MACKIEWICZ P., MACKIEWICZ D., NOWICKA A., DUDKIEWICZ M., DUDEK M.R., CEBRAT S. (2001b). High correlation between the turnover of nucleotides under mutational pressure and the DNA composition. BMC Evol. Biol. 1(1): 13.

KUMAR S., TAMURA K., NEI M. (1993). MEGA: Molecular Evolutionary Genetics Analysis. Pennsylvania State University, University Park, PA.

KUNST F., OGASAWARA N., MOSZER I., ALBERTINI A.M., ALLONI G., AZEVEDO V., BERTERO M.G., BESSIERES P., BOLOTIN A., BORCHERT S. et al. (1997). The complete genome sequence of the gram-positive bacterium *Bacillus subtilis*. Nature 390: 249-256.

LAFAY B., LLOYD A.T., MCLEAN M.J., DEVINE K.M., SHARP P.M., WOLFE K.H. (1999). Proteome composition and codon usage in spirochaetes: species-specific and DNA strand-specific mutational biases. Nucleic Acids Res. 27: 1642-1649.

LI W.-H., GOJOBORI T., NEI M. (1981). Pseudogenes as a paradigm of neutral evolution. Nature 292: 237-239.

LOBRY J.R. (1996). Asymmetric substitution patterns in the two DNA strands of bacteria. Mol. Biol. Evol. 13: 660-665.

LOPEZ P., PHILIPPE H. (2001). Composition strand asymmetries in prokaryotic genomes: mutational bias and biased gene orientation. C. R. Acad. Sci. III, 324: 201-208.

LYNCH M., CONERY J.S. (2000). The evolutionary fate and consequences of duplicate genes. Science 290: 1151-1155.

MACKIEWICZ P., GIERLIK A., KOWALCZUK M., DUDEK M.R., CEBRAT S. (1999a). Asymmetry of nucleotide composition of prokaryotic chromosomes. J. Appl. Genet. 40: 1-14.

MACKIEWICZ P., GIERLIK A., KOWALCZUK M., DUDEK M.R., CEBRAT S. (1999b). How does replication-associated mutational pressure influence amino acid composition of proteins? Genome Res. 9: 409-416.

MACKIEWICZ P., KOWALCZUK M., MACKIEWICZ D., NOWICKA A., DUDKIEWICZ M., LASZKIEWICZ A., DUDEK M.R., CEBRAT S. (2002). How many protein-coding genes are there in the *Saccharomyces cerevisiae* genome? Yeast 19: 619-629.

MACKIEWICZ P., SZCZEPANIK D., GIERLIK A., KOWALCZUK M., NOWICKA A., DUDKIEWICZ M., DUDEK M.R., CEBRAT S. (2001). The differential killing of genes by inversions in prokaryotic genomes. J. Mol. Evol. 53: 615-621.

MCINERNEY J.O. (1998). Replicational and transcriptional selection on codon usage in *Borrelia burgdorferi*. Proc. Natl. Acad. Sci. USA 95: 10698-10703.

MCLEAN M.J., WOLFE K.H., DEVINE K.M. (1998). Base composition skews, replication orientation, and gene orientation in 12 prokaryote genomes. J. Mol. Evol. 47: 691-696.

MIGHELL A.J., SMITH N.R., ROBINSON P.A., MARKHAM A.F. (2000). Vertebrate pseudogenes. FEBS Lett. 468: 109-114.

MIRA A., OCHMAN H., MORAN N.A. (2001). Deletional bias and the evolution of bacterial genomes. Trends Genet. 17: 589-596.

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

OCHMAN H. (2002). Distinguishing the ORFs from the ELFs: short bacterial genes and the annotation of genomes. Trends Genet. 18: 335-337.

OGATA H., AUDIC S., RENESTO-AUDIFFREN P., FOURNIER P.E., BARBE V., SAMSON D., ROUX V., COSSART P., WEISSENBACH J., CLAVERIE J.M., RAOULT D. (2001). Mechanisms of evolution in *Rickettsia conorii* and *R. prowazekii*. Science 293: 2093-2098.

PARKHILL J., WREN B.W., THOMSON N.R., TITBALL R.W., HOLDEN M.T., PRENTICE M.B., SEBAIHIA M., JAMES K.D., CHURCHER C., MUNGALL K.L. et al. (2001). Genome sequence of *Yersinia pestis*, the causative agent of plague. Nature 413: 523-527.

PERRIERE G., LOBRY J.R., THIOULOUSE J. (1996). Correspondence discriminant analysis: a multivariate method for comparing classes of protein and nucleic acids sequences. Comput. Appl. Biosci. 12: 519-524.

PHILIPPE H, LAURENT J. (1998). How good are deep phylogenetic trees? Curr. Opin. Genet. Dev. 8: 616-623

QIAN J., LUSCOMBE N.M., GERSTEIN M. (2001). Protein family and fold occurrence in genomes: power-law behaviour and evolutionary model. J. Mol. Biol. 313: 673-681.

ROCHA E.P., DANCHIN A., VIARI A. (1999). Universal replication biases in bacteria. Mol. Microbiol. 32: 11-16.

ROCHA E.P., DANCHIN A. (2001). Ongoing evolution of strand composition in bacterial genomes. Mol. Biol. Evol. 18: 1789-1799.

ROMERO H., ZAVALA A., MUSTO H. (2000). Codon usage in *Chlamydia trachomatis* is the result of strand-specific mutational biases and a complex pattern of selective forces. Nucleic Acids Res. 28: 2084-2090.

SCHMIDT H.A., STRIMMER K., VINGRON M., von HAESELER A. (2002). TREE-PUZZLE: maximum likelihood phylogenetic analysis using quartets and parallel computing. Bioinformatics 18: 502-504.

SHARP P.M., LI W.-H. (1987). The rate of synonymous substitution in enterobacterial genes is inversely related to codon usage bias. Mol. Biol. Evol. 4: 222-230.

SŁONIMSKI P.P., MOSSE M.O., GOLIK P., HENAUT A., DIAZ Y., RISLER J.L., COMET J.P. AUDE J.C., WO NIAK A., GLEMET E. et al. (1998). The first laws of genomics. Microb. Comp. Genomics 3: 46.

SOKAL R., ROHLF F.J. (1995). Biometry. Freeman, New York.

SZCZEPANIK D., MACKIEWICZ P., KOWALCZUK M., GIERLIK A., NOWICKA A., DUDEK M.R., CEBRAT S. (2001). Evolution rates of genes on leading and lagging DNA strands. J. Mol. Evol. 52: 426-433.

TAMURA K., NEI M. (1993). Estimation of the number of nucleotide substitutions in the control region of mitochondrial DNA in humans and chimpanzees. Mol. Biol. Evol. 10: 512-526.

TATUSOV R.L., NATALE D.A., GARKAVTSEV I.V., TATUSOVA T.A., SHANKAVARAM U.T., RAO B.S., KIRYUTIN B., GALPERIN M.Y., FEDOROVA N.D.,

KOONIN E.V. (2001). The COG database: new developments in plogenetic classification of proteins from complete genomes. Nucleic Acids Res. 29: 22-28.

THOMPSON J.D., HIGGINS D.G., GIBSON T.J. (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 E.R., COLLINS R.A. (2000a). Replication orientation affects the rate and direction of bacterial gene evolution. J. Mol. Evol. 51: 459-463.

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