Acta Microbiologica Polonica
2003, Vol. 52, No 3, 245–261

Rearrangements between Differently Replicating DNA Strands in Asymmetric Bacterial Genomes

DOROTA MACKIEWICZ[1], PAWEŁ MACKIEWICZ[1], MARIA KOWALCZUK[1],
MAŁGORZATA DUDKIEWICZ[1], MIROSŁAW R. DUDEK[2] and STANISŁAW CEBRAT[1*]

[1] Institute of Genetics and Microbiology, Wrocław University,
ul. Przybyszewskiego 63/77, 51-148 Wrocław, Poland;
[2] Institute of Physics, University of Zielona Góra,
ul. Wojska Polskiego 69, 65-246 Zielona Góra, Poland,

Received 21 May 2003

Abstract

Many bacterial genomes are under asymmetric mutational pressure which introduces compositional asymmetry into DNA molecule resulting in many biases in coding structure of chromosomes. One of the processes affected by the asymmetry is translocation changing the position of the coding sequence on chromosome in respect to the orientation on the leading and lagging DNA strand. When analysing sets of paralogs in 50 genomes, we found that the number of observed genes which switched their positions on DNA strand is lowest for genomes with the highest DNA asymmetry. However, the number of orthologs which changed DNA strand increases with the phylogenetic distance between the compared genomes. Nevertheless, there is a fraction of coding sequences that stay on the leading strand in all analysed genomes, whereas there are no sequences that stay always on the lagging strand. Since sequences diverge very fast after switching the DNA strand, this bias in mobility of sequences is responsible, in part, for higher divergence rates among some of coding sequences located on the lagging DNA strand.

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

Introduction

Rearrangements are common in bacterial genomes (Mushegian and Koonin, 1996; Tatusov *et al.*, 1996; Kolsto, 1997; Watanabe *et al.*, 1997; Bellgard *et al.*, 1999; Itoh *et al.*, 1999; Hughes, 2001) but this phenomenon has not been analysed with respect to leading/lagging strand asymmetry of bacterial chromosomes which seems to be a characteristic (if not universal) feature of these genomes (*e.g.*

[*] Address for correspondence: cebrat@microb.uni.wroc.pl; tel. + 48-71-3756-303; fax: + 48-71-3252-151

Lobry, 1996; Freeman *et al.*, 1998; Grigoriev, 1998; McLean *et al.*, 1998; Mackiewicz *et al.*, 1999a; Rocha *et al.*, 1999; Tillier and Collins, 2000a; see for review: Francino and Ochman, 1997; Mrazek and Karlin, 1998; Frank and Lobry, 1999; Kowalczuk *et al.*, 2001a). Rearrangements of genes in bacterial chromosomes follow very specific rules. In very closely related genomes, many observed rearrangements are symmetric with respect to the origin or terminus of replication (Eisen *et al.*, 2000; Read *et al.*, 2000; Tillier and Collins, 2000b; Suyama and Bork, 2001). Tillier and Collins (2000b) claim that such rearrangements are a result of higher frequency of recombination events at the replication forks which might be recombination hot spots. Another explanation involves the role of selection, and is supported by many genetic and experimental analyses (Schmid and Roth, 1983; Mahan and Roth, 1988; 1991; Rebollo *et al.*, 1988; Segall *et al.*, 1988; Segal and Roth, 1989; Francois *et al.*, 1990; Liu and Sanderson, 1995; 1996; Sanderson and Liu, 1998; Alokam *et al.*, 2002). The distance from the origin of replication determines copy number of a gene (dosage effect). Thus, genes should be located in optimal distances from the origin, according to their required expression level. There is also a trend to keep the same size of both replichores which ensures the shortest time of chromosome replication. Furthermore, since inversions of sequences resulting in switching the position of the coding sequence with respect to leading/lagging role of DNA strand is connected with a higher mutational pressure (Tillier and Collins, 2000c; Rocha and Danchin, 2001; Szczepanik *et al.*, 2001), there could be a higher probability that such a sequence will be eliminated by selection (Mackiewicz *et al.*, 2001a). (In the terminology, a coding sequence is supposed to be positioned on the leading strand if its sense strand is on the leading DNA strand, respectively the same for the lagging DNA strand). An inversion of a chromosome fragment which encompasses the origin or the terminus of replication does not change the positions of sequences in respect to the leading/lagging role of the DNA strand (Mackiewicz *et al.*, 2001b). This could lead to a bias in the observed rearrangements. Actually, experimental analyses have shown that permissive (viable) chromosome rearrangements include the origin or terminus of replication (Schmid and Roth, 1983; Mahan and Roth, 1991; Alokam *et al.*, 2002). Nevertheless, this feature of keeping the same distance from the origin of replication disappears very fast with phylogenetic distance between analysed genomes which leaves an impression that there is no structural correlation between chromosomes of distant genomes (Eisen *et al.*, 2000; Tillier and Collins, 2000b). On the other hand, there are some other phenomena, which could introduce some correlation or structural bias across genomes even at higher phylogenetic distances. Such a phenomenon is a differentiated mutational pressure for coding sequences located on the leading and the lagging strands (Tillier and Collins, 2000c; Rocha and Danchin, 2001; Szczepanik *et al.*, 2001). There appears to be some preference in the accumulation of translocated coding sequences from the lagging to the leading strand rather than in the opposite direction (McInerney, 1998; Mackiewicz *et al.*, 2001a). Again, mechanisms of selection are blamed for this bias rather than bias in frequency of translocations themselves. A significant surplus of genes on the leading strand has been observed in many genomes (Brewer, 1988; Fraser *et al.*, 1995; Kunst *et al.*, 1997; Freeman *et al.*, 1998; McLean *et al.*, 1998). Knowing that the divergence rate of coding sequences depends on their location on the leading/lagging DNA strand (Szczepanik *et al.*, 2001), we should expect also a correlation between the function of genes and their position on chromosome as well as differentiated frequency of switching the position of genes lying on the two DNA strands.

One of the main mechanisms of genome evolution is gene duplication, which enables further independent evolution of the structure and function of the two copies (Ohno, 1970). These copies can be seen in genomes as paralogs – homologous sequences occurring in the same genome (Fitch, 1970). It was found that both, duplication and elimination of paralogs should be ruled by some strict mechanisms, since the number of paralogs follows a very specific numerical law (Huynen and Nimwegen, 1998; Slonimski *et al.*, 1998; Qian *et al.*, 2001). What we observe is a final result of duplication itself and the paralogs elimination. Duplication of sequences could be connected with a transfer of a new copy into the other DNA strand (inversion) or the copy could stay at the same strand. The mutation rate in sequences after inversion is higher, thus there should be a higher elimination rate of inverted copies. We have already shown that it is true (Mackiewicz *et al.*, 1999a). Genes which have switched DNA strand accommodate very quickly to a new mutational pressure and, in respect to their nucleotide composition, become similar to genes of the new strand (Lafay *et al.*, 1999; Tillier and Collins, 2000c; Rocha and Danchin, 2001).

In this paper we present the results of analysis of fully sequenced bacterial genomes which revealed asymmetry in frequency of translocations (viable inversions) of genes lying on the leading and the lagging DNA strands and we have shown how this affects the divergence rate of genes classified according to the criteria of their mobility.

Experimental

Materials and Methods

Data for analysis. 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 nucleotide compositional bias of differently replicating DNA strands (Mackiewicz *et al.*, 1999b, see also: http://smorfland.microb.uni.wroc.pl). The asymmetry of the genomes was measured by the absolute value of the difference between the GC3 skews of the genes in the leading strand and the ones in the lagging strand:

ΔGC3 skew = $|(G_d - C_d)/(G_d + C_d) - (G_g - C_g)/(G_g + C_g)|$

where: G_d and C_d – numbers of guanine and cytosine in the third codon positions of the leading strand genes; G_g and C_g – numbers of guanine and cytosine in the third codon positions of the lagging strand genes. The AT skew and GC skew values proved to be good parameters describing asymmetry of DNA strands (Lobry, 1996).

Paralogs for 50 genomes (listed in Table I) showing leading/lagging strand asymmetry were extracted from the TIGR database (http://www.tigr.org). In the analysis only paralogs with minimum 50% identity were chosen.

Classification of genes to orthologous groups and their amino acid sequences were extracted from Clusters of Orthologous Groups (COGs) downloaded from ftp://www.ncbi.nlm.nih.gov/pub/COG in September 2001. COGs contain protein sequences which are supposed to have evolved from one ancestral protein (Koonin *et al.*, 1998; Tatusov *et al.*, 2001). In the analyses only the best matches for each ortholog (the closest orthologs) have been chosen.

Analyses of all orthologous sequences have been done on the two sets of bacterial genomes showing evident compositional asymmetry between leading and lagging strands.

- 7 genomes belonging to γ-subdivision of Proteobacteria group compared with each other: *E. coli* K12-MG1655 (EcK), *E. coli* O157:H7 EDL933 (EcE), *H. influenzae* (Hi), *P. multocida* (Pm), *P. aeruginosa* (Pa), *V. cholerae* (Vc), *X. fastidiosa* (Xf);
- 14 genomes compared with *E. coli* O157:H7 EDL933 (EcE): *E. coli* K12-MG1655 (EcK), *V. cholerae* (Vc), *P. multocida* (Pm), *P. aeruginosa* (Pa), *X. fastidiosa* (Xf), *N. meningitidis* MC58 (Nm), *B. subtilis* (Bs), *R. prowazekii* (Rp), *M. tuberculosis* H37Rv (Mt), *C. jejuni* (Cj), *T. pallidum* (Tp), *H. pylori* 26695 (Hp), *C. pneumoniae* CWL029 (Cp), *P. horikoshii* (Ph).

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. Similarly, from the set of 14 genomes compared with *E. coli* EDL933, the 14 sets of 233 orthologs present in all the genomes, being the "best hits" for *E. coli* EDL933 sequences, were extracted.

For each pair of genomes, orthologs and paralogs were classified into three groups according to their strand location: pairs of sequences lying on the leading strands, pairs of sequences lying on lagging strands, and pairs of sequences of which one is lying on the leading and the other on the lagging strand. For each case fractions of the three groups of sequences have been counted.

Phylogenetic analysis. The amino acid sequences of each COG were aligned by the CLUSTAL W 1.8 v. program (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 using the WAG model of amino acid substitution (Whelan and Goldman, 2001) as implemented in the TREE-PUZZLE program version 5.0 (Schmidt *et al.*, 2002). The analyses of divergence of the three groups of orthologs were shown for the sets of 1521 orthologs present in all 7 γ-Proteobacteria genomes.

For each of the three groups of orthologs a mean value of the evolutionary distances was calculated. Nonparametric analyses by Mann-Whitney U, Kolmogorov-Smirnov and ANOVA Kruskal-Wallis tests (Sokal and Rohlf, 1995) were carried out to assess statistical significance of differences between these groups.

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 Tamura-Nei model of nucleotide substitutions (Tamura and Nei, 1993).

Results and Discussion

In highly asymmetric genomes, the mutational pressure after inversion should be relatively higher than for genomes with low asymmetry – there are stronger differences in substitution rates for the leading and lagging DNA strands in the asymmetric genomes (Kowalczuk *et al.*, 2001b; Rocha and Danchin, 2001). Thus, we have anticipated and found a negative correlation between the chromosome asymmetry and the frequency of occurring paralogs in the trans-positions in the genome (one paralog on the leading strand, the other one on the lagging strand – we call these sequences "trans-paralogs"). In Table I we show data for each analysed genome and in Fig. 1 we show the relation between the fraction of trans-paralogs in the genome and the asymmetry of chromosomes measured by ΔGC3 skew. The observed negative correlation (Spearman correlation coefficient, $r = -0.715$) is statistically significant with high confidence ($p = 5.6 \times 10^{-9}$). There are two possible explanations for the observed negative correlation. One, assuming a higher mutation rate and in consequence higher elimination rate of gene copies translocated to the other DNA strand in highly asymmetric genomes. The second, to us less plausible, refers to the influence of frequency of rearrangements on the maintenance of chromosomal asymmetry. If a global frequency of rearrangements in a genome is low, it does not disturb chromosomal asymmetry established by the mutational pressure. On the contrary, high frequency of rearrangements should diminish this asymmetry.

We have performed a pairwise analysis of orthologs found in compared genomes belonging to γ-Proteobacteria. For each pair of genomes, the orthologs were divided into three groups: i/ pairs of orthologs which are in both compared genomes on the leading strand, ii/ pairs of orthologs which are in both genomes on the lagging strand and iii/ pairs of orthologs of which one is located on the leading and the second on the

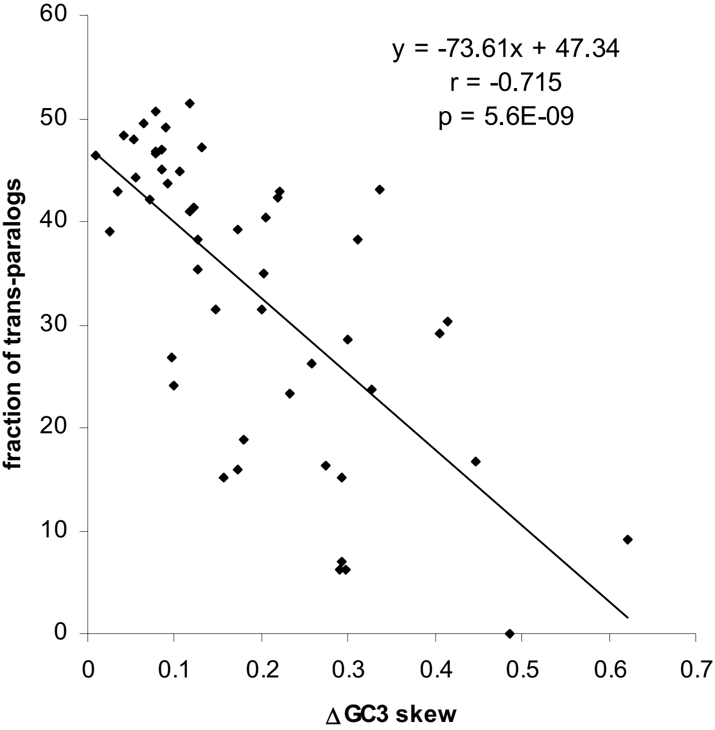

Fig. 1. Relation between the fraction of trans-paralogs (one paralog on the leading strand, the other one on the lagging strand) in 50 analysed genomes and the asymmetry of chromosomes measured by ΔGC3 skew. Spearman correlation coefficient (r) and its statistical significance (p) are shown.

Table I
Number of all paralogs, the fraction of trans-paralogs and DGC3 skew for 50 analysed genomes

genome	number of all paralogs	fraction of trans-paralogs	ΔGC3 skew
Agrobacterium tumefaciens C58 Cereon	1096	46.8	0.08
Agrobacterium tumefaciens C58 Uwash	1117	46.6	0.08
Bacillus halodurans C-125	2421	39.2	0.17
Bacillus subtilis 168	558	31.4	0.15
Borrelia burgdorferi B31	11	9.1	0.62
Brucella melitensis 16M	314	38.2	0.13
Campylobacter jejuni NCTC 11168	79	30.4	0.41
Caulobacter crescentus CB15	511	44.2	0.05
Chlamydia muridarum Nigg	19	0.0	0.49
Chlamydia pneumoniae AR39	111	6.3	0.29
Chlamydia pneumoniae CWL029	112	6.3	0.30
Chlamydia pneumoniae J138	100	7.0	0.29
Chlamydia trachomatis serovar D	6	16.7	0.45
Clostridium perfringens 13	217	29.0	0.41
Deinococcus radiodurans R1	282	46.5	0.01
Escherichia coli O157:H7 EDL933	3604	26.9	0.10
Escherichia coli K12-MG1655	919	47.0	0.09
Escherichia coli VT2-Sakai	4020	24.1	0.10
Haemophilus influenzae KW20	73	15.1	0.16
Helicobacter pylori 26695	198	51.5	0.12
Helicobacter pylori J99	109	41.3	0.12
Lactococcus lactis IL1403	811	42.9	0.22
Listeria innocua CLIP 11262	349	18.9	0.18
Listeria monocytogenes EGD-e	255	31.4	0.20
Mesorhizobium loti MAFF303099	1414	42.9	0.04
Mycobacterium leprae TN	121	47.1	0.13
Mycobacterium tuberculosis CDC1551	2417	43.7	0.09
Mycobacterium tuberculosis H37Rv	2279	45.1	0.08
Neisseria meningitidis MC58	595	35.0	0.20
Neisseria meningitidis Z2491	874	42.3	0.22
Pasteurella multocida PM70	86	23.3	0.23
Pseudomonas aeruginosa PAO1	786	44.8	0.11
Pyrococcus abyssi GE5	118	48.3	0.04
Pyrococcus horikoshii shinkaj OT3	147	42.2	0.07
Ralstonia solanacearum GMI1000	784	50.8	0.08
Rickettsia conorii Malish 7	478	40.4	0.21

Table I continued

genome	number of all paralogs	fraction of trans-paralogs	ΔGC3 skew
Salmonella enterica Typhi CT18	679	35.3	0.13
Salmonella typhimurium LT2 SGSC1412	1280	41.0	0.12
Sinorhizobium meliloti 1021	460	48.0	0.05
Staphylococcus aureus Mu50	228	16.2	0.28
Staphylococcus aureus N315	482	15.1	0.29
Streptococcus pneumoniae R6	759	28.5	0.30
Streptococcus pneumoniae TIGR4	479	38.2	0.31
Streptococcus pyogenes SF370 M1	130	26.2	0.26
Thermoplasma acidophilum DSM 1728	41	39.0	0.03
Thermotoga maritima MSB8	216	49.5	0.06
Treponema pallidum Nichols	72	43.1	0.34
Vibrio cholerae El Tor N16961	868	16.0	0.17
Xylella fastidiosa 9a5c	779	23.7	0.33
Yersinia pestis CO92	8551	49.1	0.09

The set of paralogs (with minimum 50 % identity) was extracted from TIGR database.

lagging strand. If we assume that there is no bias in the frequency of inversions of genes located on the leading and on the lagging DNA strands, we should expect that the fractions of orthologs staying at the same strand in both genomes of the compared pair would decrease with the phylogenetic distance between genomes but the decrease should be proportional to the initial values on the two strands. The results of analyses do not follow these expected rules.

For each pair of compared genomes we have plotted (Fig. 2) the fractions of the three groups of orthologs against the evolutionary distance measured by divergence of 16S rRNA genes between the two compared genomes. The fraction of orthologs lying on the same strand decreases with evolutionary distance while fraction of orthologs which have switched their strands increases rapidly with divergence and become saturated for long evolutionary distances. The same results we have obtained for similar analysis when we compared the *E. coli* EDL933 genome with 14 other genomes belonging to different taxonomic groups (Fig. 3). Even at a short distance (up 0.22 of divergence of 16S rRNA), the total fraction of sequences which switched their strand reaches almost 50%. But there is a very biased input of sequences from the leading and the lagging DNA strands into this fraction. While the fraction of sequences which stay at the leading strands in both compared genomes drops to about 70% of the initial value in the most distant pair, the relative numbers for the lagging strand are up to 40%. These results suggest that the sequences lying on the lagging strand are much more prone to inversions than the sequences lying on the leading strand.

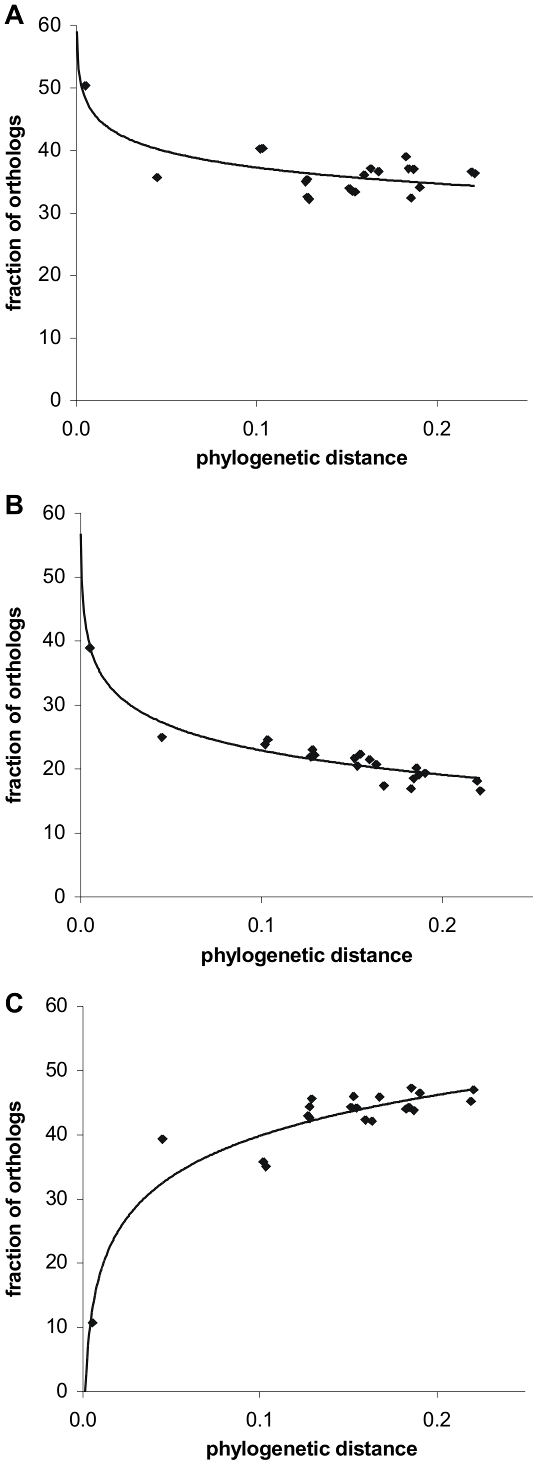

Fig. 2. Relation between the fractions of orthologs and the phylogenetic distance measured by 16S rRNA performed for three groups of orthologs: lying on the leading strand (A), lying on the lagging strand (B) and which changed DNA strand (C).

Data obtained from pairwise comparison of 7 genomes belonging to g- Proteobacteria.

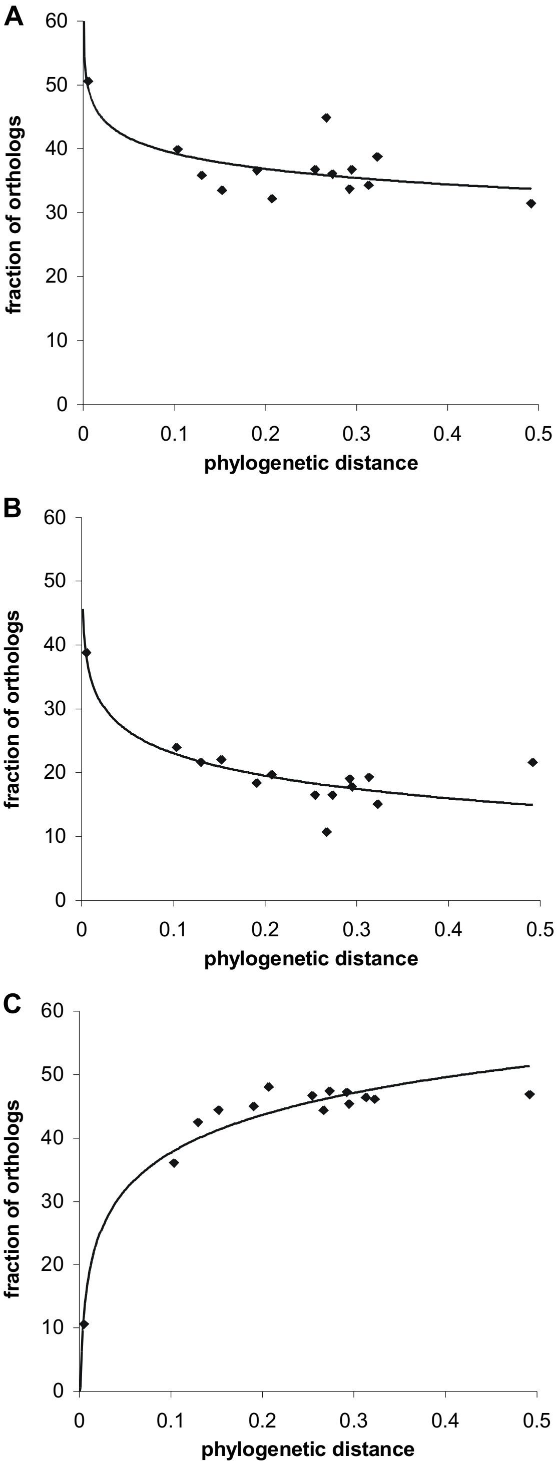

Fig. 3. Relation between the fractions of orthologs and the phylogenetic distance measured by 16S rRNA performed for three groups of orthologs: lying on the leading strand (A), lying on the lagging strand (B) and which changed DNA strand (C).

Data obtained from comparison of the *E. coli* EDL933 genome with 14 other genomes belonging to different taxonomic groups.

This observation implies also that there are some sequences which "are used to" staying on the leading DNA strand and they have lower probability of being inverted than sequences which "are used to" staying on the lagging strand. As a consequence, the set of coding sequences found on the leading strand should be not uniform. It should consist of a set of sequences which permanently or preferentially stay on the leading strand and a set of mobile sequences which are only transiently transferred from the lagging strand. To test this hypothesis we analysed the sets of 233 orthologs represented in all 15 genomes. In the first step we compared the most closely related genomes in the analysed set – two *E. coli* strains – and we counted the fractions of orthologs which stayed at the same DNA strands (leading or lagging) and the fraction of orthologs which switched their strands. In the next step we added to the comparison the third genome (the closest to the *E. coli* EDL933 genome according to the 16S rRNA phylogenetic distance) and again counted sequences which stayed at the same DNA strand in all the three genomes and sequences which switched their strand at least in one genome and so on, adding new, more distant genome to the analysed group. In Fig. 4, in the diagram, we have presented the results of analysis; values on y-axis correspond to the fraction of sequences of a given group of orthologs, while at the bottom the name of a new genome added to the comparison is shown. The fraction

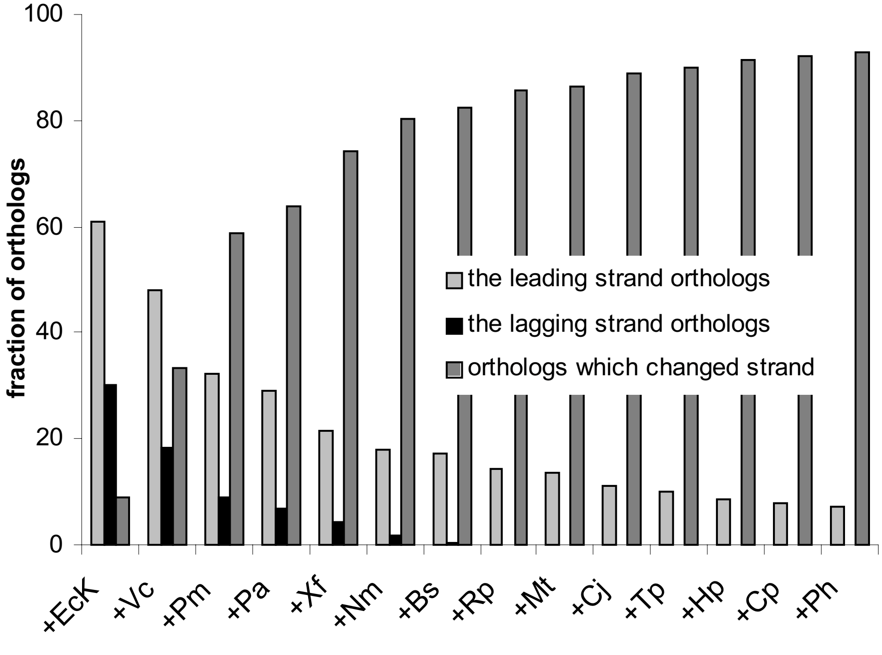

Fig. 4. The fractions of three groups of orthologs counted for the comparisons of *E. coli* EDL933 with successively added genomes to the comparison.

The group of the leading strand orthologs contains sequences which stay on the leading strand in all analysed genomes in a given comparison. Analogously for the lagging strand orthologs. The third group of orthologs includes sequences which switched their strand at least in one genome in a given comparison. Data were obtained for the sets 233 orthologs present in all 15 genomes. For genomes name abbreviations see Materials and Methods.

Table II
Orthologs found in all 15 analysed genomes
on the leading strand

COG number	description
COG0051	Ribosomal protein S10
COG0087	Ribosomal protein L3
COG0088	Ribosomal protein L4
COG0090	Ribosomal protein L2
COG0091	Ribosomal protein L22
COG0092	Ribosomal protein S3
COG0093	Ribosomal protein L14
COG0094	Ribosomal protein L5
COG0096	Ribosomal protein S8
COG0097	Ribosomal protein L6
COG0098	Ribosomal protein S5
COG0185	Ribosomal protein S19
COG0186	Ribosomal protein S17
COG0197	Ribosomal protein L16/L10E
COG0198	Ribosomal protein L24
COG0200	Ribosomal protein L15
COG0256	Ribosomal protein L18

of sequences which stay in all analysed genomes on the lagging strand drops very fast and after adding the eighth genome it reaches zero, which means that there are no orthologous coding sequences located on the lagging strands in all compared genomes. For this group of compared genomes, there are still some orthologs which stay on the leading strand in all the genomes and this fraction seems to approximate asymptotically about 7% of all compared coding sequences, even after adding the most distant genome belonging to Archaea. These orthologs code for ribosomal proteins commonly considered highly conserved (Table II). The position of these genes on the leading strand seems to be conserved even across the two kingdoms (Bacteria and Archaea). It was observed that their operons are well preserved even in divergent species (Watanabe *et al.*, 1997; Itoh *et al.*, 1999; Nikolaichik and Donachie, 2000; Tamames, 2001). Moreover, it was found that ribosomal genes are preferentially located in many genomes on the leading strand (McLean *et al.* 1998) probably (what is important for highly expressed genes) to avoid head-on collisions between replication and transcription complexes (Brewer, 1988; French, 1992).

In the next studies we have analysed the divergence measured by the mean number of amino acid substitutions per site in groups of sequences classified according to their mobility between differently replicating DNA strands. Analyses were performed with the sets of 1521 orthologs present in all 7 genomes belonging to γ-Proteobacteria. We compared the *E. coli* EDL933 genome with six other genomes.

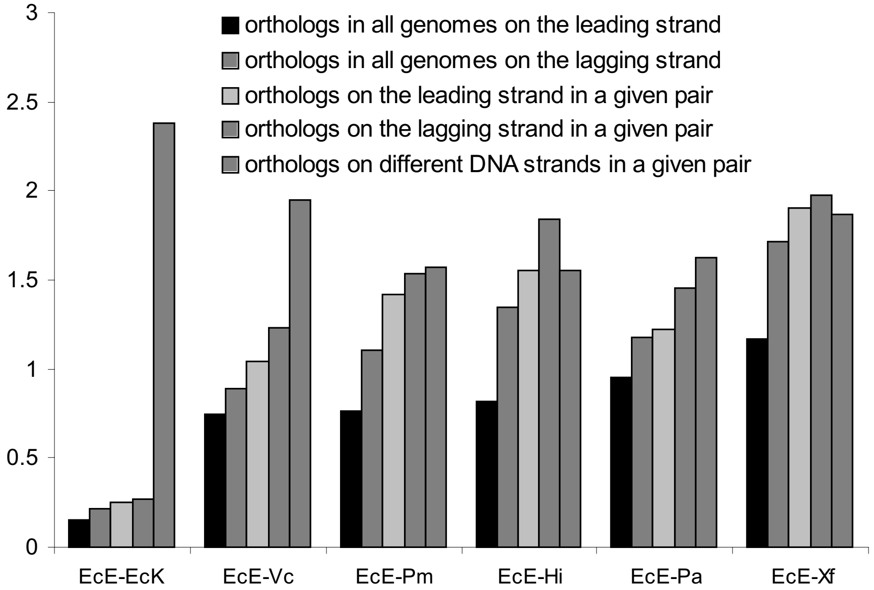

Fig. 5. The divergence measured by the mean number of amino acid substitutions per site according to WAG model (Whelan and Goldman, 2001) in five groups of sequences classified according to their mobility between differently replicating DNA strands for comparisons of *E. coli* EDL933 with other genomes. Analyses were performed with the sets of 1521 orthologs present in all 7 genomes belonging to γ-Proteobacteria. For genomes name abbreviations see Materials and Methods.

We divided all orthologs into five sets: 1 – genes staying in all analysed genomes on the leading strand, 2 – genes staying in all analysed genomes on the lagging strand, 3 – genes which are located in *E. coli* EDL933 and in the compared genome on the leading strand but can be found in at least one of the other genomes of γ-Proteobacteria on the lagging strand, 4 – genes which are located in *E. coli* EDL933 and in the compared genome on the lagging strand but can be found in at least one of the other genomes on the leading strand and, 5 – sequences which are located on different DNA strands in the compared genomes. The divergence values between genes of the *E. coli* EDL933 genome and other genomes of γ-Proteobacteria are shown in Fig. 5. We have found that there are statistically significant differences in the relative divergence between genes classified according to their position and mobility. The differences between set 1 and set 5 are statistically significant (with $p < 0.01$) for all comparisons. It is clear that the divergence of the orthologs which switched strand (set 5) is especially high for the closest genomes, which was already reported (Tillier and Collins, 2000c; Szczepanik *et al.*, 2001; Rocha and Danchin, 2001) and decreases for pairs of distant genomes. Differences in divergence between set 5 and all other sets are statistically significant (with $p<0.01$) for pairs: EcE-EcK, EcE-Vc and EcE-Pa.

In all compared pairs of genomes the lowest divergence is observed for the orthologs which permanently stay at the leading strand and do not change their strand even at long evolutionary distances (set 1). If we eliminate this set of conserved genes from the set of all orthologs found on the leading strand (receiving set 3), the rest still seems to be less prone to accumulate substitutions than the genes from the lagging strand. However, we have found only one statistically significant difference in divergence (5.6% of all comparisons) when we compared sets 2, 3 and 4 with each other for all pairs of genomes. Furthermore, the divergence in these three sets is significantly different (with $p < 0.01$) when analysed by the ANOVA Kruskal-Wallis test only for one pair EcE-Pa. It indicates that these three sets form rather uniform group.

Conclusions

The observed rearrangements in bacterial chromosomes are not random. Mutational pressure, responsible for the observed asymmetry in DNA composition, affects especially the copies of genes translocated to other DNA strand. According to the mobility (frequency of translocations between leading and lagging strand) it is possible to classify genes into two groups: highly conserved genes permanently or preferentially lying on the leading strand and genes switching their position between the leading and lagging DNA strands.

Acknowledgments. The work was supported by the grant number 1016/S/IMi/03. M.K. was supported by Foundation for Polish Science.

Literature

Alokam S., S.L. Liu, K. Said and K.E. Sanderson. 2002. Inversions over the terminus region in *Salmonella* and *Escherichia coli*: IS200s as the sites of homologous recombination inverting the chromosome of *Salmonella enterica* serovar typhi. *J. Bacteriol.* **184**: 6190–6197.

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

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

Eisen J.A., J.F. Heidelberg, O. White and S.L. Salzberg. 2000. Evidence for symmetric chromosomal inversions around the replication origin in bacteria. *Genome Biol.* **1**(6): research0011.

Fitch W.M. 1970. Distinguishing homologous from analogous proteins. *Sys Zool.* **19**: 99–113.

Francino M.P. and H. Ochman. 1997. Strand asymmetries in DNA evolution. *Trends Genet.* **13**: 240–245.

Francois V., J. Louarn, J. Patte, J.E. Rebollo and J.M. Louarn. 1990. Constraints in chromosomal inversions in *Escherichia coli* are not explained by replication pausing at inverted terminator-like sequences. *Mol. Microbiol.* **4**: 537–542.

Frank A.C. and J.R. Lobry. 1999. Asymmetric substitution patterns: a review of possible underlying mutational or selective mechanisms. *Gene* **238**: 65–77.

Fraser C.M., J.D. Gocayne, O. White, M.D. Adams, R.A. Clayton, R.D. Fleischmann, C.J. Bult, A.R. Kerlavage, G.G. Sutton, J.M. Kelley, *et al.* 1995. The minimal gene complement of *Mycoplasma genitalium*. *Science* **270**: 397–403.

Freeman J.M., T.N. Plasterer, T.F. Smith and S.C. Mohr. 1998. Patterns of genome organization in bacteria. *Science* **279**: 1827.

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

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

Hughes D. 2000. Evaluating genome dynamics: the constraints on rearrangements within bacterial genomes. *Genome Biol.* **1**(6):reviews0006.

Huynen M.A. and E. van Nimwegen. 1998. The frequency distribution of gene family sizes in complete genomes. *Mol. Biol. Evol.* **15**: 583–589.

Itoh T., K. Takemoto, H. Mori and T. Gojobori. 1999. Evolutionary Instability of Operon Structures Disclosed by Sequence Comparisons of Complete Microbial Genomes. *Mol. Biol. Evol.* **16**: 332–346.

Kolsto A.B. 1997. Dynamic bacterial genome organization. *Mol. Microbiol.* **24**: 241–248.

Koonin E.V., R.L. Tatusov and M.Y. Galperin. 1998. Beyond complete genomes: from sequence to structure and function. *Curr. Opin. Struct. Biol.* **8**: 355–363.

Kowalczuk M., P. Mackiewicz, D. Mackiewicz, A. Nowicka, M. Dudkiewicz, M.R. Dudek and S. Cebrat. 2001a. DNA asymmetry and the replicational mutational pressure. *J. Appl. Genet.* **42**: 553–577.

Kowalczuk M., P. Mackiewicz, D. Mackiewicz, A. Nowicka, M. Dudkiewicz, M.R. Dudek and S. Cebrat. 2001b. High correlation between the turnover of nucleotides under mutational pressure and the DNA composition. *BMC Evol. Biol.* **1**(1):13.

Kumar S., K. Tamura and M. Nei. 1993. MEGA: Molecular Evolutionary Genetics Analysis. Pennsylvania State University, University Park, PA.

Kunst F., N. Ogasawara, I. Moszer, A.M. Albertini, G. Alloni, V. Azevedo, M.G. Bertero, P. Bessieres, A. Bolotin, S. Borchert, *et al.* 1997. The complete genome sequence of the gram-positive bacterium *Bacillus subtilis*. *Nature* **390**: 249–256.

Lafay B., A.T. Lloyd, M.J. McLean, K.M. Devine, P.M. Sharp and K.H. Wolfe. 1999. Proteome composition and codon usage in spirochaetes: species-specific and DNA strand-specific mutational biases. *Nucleic Acids Res.* **27**: 1642–1649.

Liu S.L. and K.E. Sanderson. 1995. Rearrangements in the genome of the bacterium *Salmonella typhi*. *Proc. Natl. Acad. Sci. USA* **92**: 1018–1022.

Liu S.L. and K.E. Sanderson. 1996. Highly plastic chromosomal organization in *Salmonella typhi. Proc. Natl. Acad. Sci. USA* **93**: 10303–10308.

Lobry J.R. 1996. Asymmetric substitution patterns in the two DNA strands of bacteria. *Mol. Biol. Evol.* **13**: 660–665.

Mackiewicz P., A. Gierlik, M. Kowalczuk, M.R. Dudek and S. Cebrat. 1999a. Asymmetry of nucleotide composition of prokaryotic chromosomes. *J. Appl. Genet.* **40**: 1–14.

Mackiewicz P., A. Gierlik, M. Kowalczuk, M.R. Dudek and S. Cebrat. 1999b. How does replication-associated mutational pressure influence amino acid composition of proteins? *Genome Res.* **9**: 409–416.

Mackiewicz P., D. Szczepanik, A. Gierlik, M. Kowalczuk, A. Nowicka, M. Dudkiewicz, M.R. Dudek and S. Cebrat. 2001a. The differential killing of genes by inversions in prokaryotic genomes. *J. Mol. Evol.* **53**: 615–621.

Mackiewicz P., D. Szczepanik, M. Kowalczuk and S. Cebrat. 2001b. Flip-flop around the origin and terminus of replication in prokaryotic genomes. *Genome Biol.* **2**(12): interactions1004.

Mahan M.J. and J.R. Roth. 1988. Reciprocality of recombination events that rearrange the chromosome. *Genetics* **120**: 23–35.

Mahan M.J. and J.R. Roth. 1991. Ability of a bacterial chromosome segment to invert is dictated by included material rather than flanking sequence. *Genetics* **129**: 1021–1032.

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., K.H. Wolfe and K.M. Devine. 1998. Base composition skews, replication orientation, and gene orientation in 12 prokaryote genomes. *J. Mol. Evol.* **47**: 691–696.

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

Mushegian A.R. and E.V. Koonin. 1996. Gene order is not conserved in bacterial evolution. *Trends Genet.* **12**: 289–290.

Nikolaichik Y.A. and W.D. Donachie. 2000. Conservation of gene order amongst cell wall and cell division genes in Eubacteria, and ribosomal genes in Eubacteria and Eukaryotic organelles. *Genetica* **108**: 1–7.

Ohno S. 1970. Evolution by gene duplication. George Allen and Unwin, London.

Qian J., N.M. Luscombe and M. Gerstein. 2001. Protein family and fold occurrence in genomes: power-law behaviour and evolutionary model. *J. Mol. Biol.* **313**: 673–681.

Read T.D., R.C. Brunham, C. Shen, S.R. Gill, J.F. Heidelberg, O. White, E.K. Hickey, J. Peterson, T. Utterback, K. Berry, *et al.* 2000. Genome sequences of *Chlamydia trachomatis* MoPn and *Chlamydia pneumoniae* AR39. *Nucleic Acids Res.* **28**: 1397–1406.

Rebollo J.E., V. Francois and J.M. Louarn. 1988. Detection and possible role of two large nondivisible zones on the *Escherichia coli* chromosome. *Proc. Natl. Acad. Sci. USA* **85**: 9391–9395.

Rocha E.P., A. Danchin and A. Viari. 1999. Universal replication biases in bacteria. *Mol. Microbiol.* **32**: 11–16.

Rocha E.P. and A. Danchin. 2001. Ongoing evolution of strand composition in bacterial genomes. *Mol. Biol. Evol.* **18**: 1789–1799.

Sanderson K.E. and S.L. Liu. 1998. Chromosomal rearrangements in enteric bacteria. *Electrophoresis* **19**: 569–572.

Schmid M.B., J.R. Roth. 1983. Selection and endpoint distribution of bacterial inversion mutations. *Genetics* **105**: 539–55.7

Schmidt H.A., K. Strimmer, M. Vingron and A. von Haeseler. 2002. TREE-PUZZLE: maximum likelihood phylogenetic analysis using quartets and parallel computing. *Bioinformatics* **18**: 502–504.

Segall A., M.J. Mahan and J.R. Roth. 1988. Rearrangement of the bacterial chromosome: forbidden inversions. *Science* **241**: 1314–1318.

Segall A.M. and J.R. Roth. 1989. Recombination between homologies in direct and inverse orientation in the chromosome of Salmonella: intervals which are nonpermissive for inversion formation. *Genetics* **122**: 737–747.

Słonimski P.P., M.O. Mosse, P. Golik, A. Henaut, Y. Diaz, J.L. Risler, J.P. Comet, J.C. Aude, A. Woźniak, E. Glemet, *et al.* 1998. The first laws of genomics. *Microb. Comp. Genomics* **3**: 46.

Sokal R.R. and F.J. Rohlf. 1995. Biometry. Freeman, New York.

Suyama M. and P. Bork. 2001. Evolution of prokaryotic gene order: genome rearrangements in closely related species. *Trends Genet.* **17**: 10–13.

Szczepanik D., P. Mackiewicz, M. Kowalczuk, A. Gierlik, A. Nowicka, M.R. Dudek and S. Cebrat. 2001. Evolution rates of genes on leading and lagging DNA strands. *J. Mol. Evol.* **52**: 426–433.

Tamames J. 2001. Evolution of gene order conservation in prokaryotes. *Genome Biol.* **2** (6): research 0020.

Tamura K. and M. Nei. 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., A.R. Mushegian, P. Bork, N.P. Brown, W.S. Hayes, M. Borodovsky, K.E. Rudd and E.V. Koonin. 1996. Metabolism and evolution of *Haemophilus influenzae* deduced from a whole-genome comparison with *Escherichia coli*. *Curr. Biol.* **6**: 279–291.

Tatusov R.L., D.A. Natale, I.V. Garkavtsev, T.A. Tatusova, U.T. Shankavaram, B.S. Rao, B. Kiryutin, M.Y. Galperin, N.D. Fedorova and E.V. Koonin. 2001. The COG database: new developments in phylogenetic classification of proteins from complete genomes. *Nucleic Acids Res.* **29**: 22–28.

Thompson J.D., D.G. Higgins and T.J. Gibson. 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. and R.A. Collins. 2000a. The contributions of replication orientation, gene direction, and signal sequences to base-composition asymmetries in bacterial genomes. *J. Mol. Evol.* **50**: 249–257.

Tillier E.R. and R.A. Collins. 2000b. Genome rearrangement by replication-directed translocation. *Nat. Genet.* **26**: 195–197.

Tillier E.R. and R.A. Collins. 2000c. Replication orientation affects the rate and direction of bacterial gene evolution. *J. Mol. Evol.* **51**: 459–463.

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

Whelan S. and N. Goldman. 2001. A general empirical model of protein evolution derived from multiple protein families using a maximum-likelihood approach. *Mol. Biol. Evol.* **18**: 691–699.