Differential gene survival under asymmetric directional mutational pressure

Paweł Mackiewicz[1], Małgorzata Dudkiewicz[1], Maria Kowalczuk[1], Dorota Mackiewicz[1], Joanna Banaszak[1], Natalia Polak[1], Kamila Smolarczyk[1], Aleksandra Nowicka[1], Mirosław R. Dudek[2], and Stanisław Cebrat[1]*

[1] Department of Genetics, Institute of Microbiology, University of Wroclaw,
ul. Przybyszewskiego 63/77, PL-54148 Wroclaw, Poland
{malgosia, pamac, nowicka, kowal, dorota, polak, smolar, cebrat}@microb.uni.wroc.pl
http://smORFland.microb.uni.wroc.pl

[2] Institute of Physics, University of Zielona Góra, ul. A. Szafrana 4a,
PL-65516 Zielona Góra, Poland
mdudek@proton.if.uz.zgora.pl

Abstract. We have simulated, using Monte Carlo methods, the survival of prokaryotic genes under directional mutational pressure. We have found that the whole pool of genes located on the leading DNA strand differs from that located on the lagging DNA strand and from the subclass of genes coding for ribosomal proteins. The best strategy for most of the non-ribosomal genes is to change the direction of the mutational pressure from time to time or to stay at their recent position. Genes coding for ribosomal proteins do not profit to such an extent from switching the directional pressure which seems to explain their extremely conserved positions on the prokaryotic chromosomes.

1 Introduction

The DNA molecule is composed of only four different kinds of nucleotides: Adenine (A), Thymine (T), Guanine (G), and Cytosine (C) repeated up to hundreds millions of times and forming two strings, oriented in opposite directions. Both the opposite directions of the strings and the possibility of synthesizing a new strand only in one direction impose a specific bias on the frequencies of occurrence of particular nucleotides in the two strands. Since A in one strand corresponds to T in the opposite position of the other strand and G corresponds to C, the next two equations are valid for the double strand molecule: [A]=[T] and [G]=[C]. This is called the complementarity rule (Parity Rule 1, PR1) and this rule is deterministic [1]. If we construct a random DNA molecule fulfilling this rule, we should expect no statistically significant differences between the numbers of A and T, and G and C in each of the single strands. This stochastic rule is called Parity Rule 2 (PR2) [2]. Any statistically significant deviation from

* To whom all correspondence should be sent.

these rules is called the DNA asymmetry. Most of the natural DNA sequences are asymmetric. There are two main mechanisms introducing DNA asymmetry: the replication-associated directional mutational pressure and the selection for protein coding sequences (see for review: [3], [4]). These two asymmetries are qualitatively and quantitatively different. The replication- associated mutational pressure generates some kind of a global asymmetry between the two strands called the leading and the lagging DNA strands ([5] – [12]). The asymmetry between the leading and the lagging DNA strands results from different patterns of nucleotide substitutions during synthesis of these two strands. On the other hand, the selection for coding sequences generates a local asymmetry between sense (coding) and anti-sense (complementary to the sense) strands of genes ([13], [15]). This asymmetry results from the coding function requirement of genes. Thus, as in the case of two chiral molecules, the two possible ways of superposition of a coding sequence on the asymmetric bacterial chromosome (sense or anti-sense of the gene replicated as the leading strand) are not equivalent. For example, if the sense strand of a gene located on the leading strand has more G than C, and C is more often substituted by other nucleotides than G on the leading strand, then inversion of this sequence, which transfers the C-rich anti-sense strand of the gene to the leading strand, would increase the mutation rate of the gene (see Fig. 1). The mutational pressure acting on the leading strand generates in the most of bacterial genomes more G than C and more T than A ([5] – [12]). Thus, a gene sequence remaining for a long time on one DNA strand tends to acquire some asymmetry characteristic for the mutational pressure while sequences occasionally inverted oscillate between the two compositional stages and their composition depends on the time which they spend on each strand and on how frequent they are translocated. In this paper we have simulated the effect of changing the mutational pressure on the gene survival. For our studies we have used parameters of the directional mutational pressure found for a real bacterial genome *Borrelia burgdorferi* and used them to mutate genes lying on the leading or lagging DNA strands in this genome. We have also analyzed a group of very conserved genes coding for ribosomal proteins.

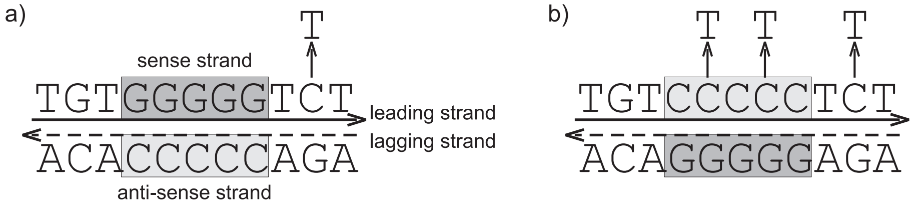

Fig. 1. The two possible ways of superposition of a coding sequence (shadowed boxes) on differently replicating DNA strands: a - location of the G-rich sense strand of a gene on the leading strand; b - the inversion of this sequence which transfers the C- rich anti-sense strand of the gene to the leading strand. Because C is more often substituted by T on the leading strand than on the lagging strand, the inversion causes an increase in the mutation rate of the gene.

2 Methods

Simulations have been performed on genes from the *B. burgdorferi* genome [16], 564 genes located on the leading DNA strand, and 286 genes from the lagging DNA strand, whose sequence and annotations were downloaded from GenBank (*ftp://ftp.ncbi.nih.gov*). We have distinguished also the third very specific, extremely conserved set of genes coding for ribosomal proteins. The replication-associated mutational pressure (RAMP) describing the nucleotide substitution frequencies has been parameterized as described by Kowalczuk et al. [17]. The matrix describing RAMP of the lagging strand is the mirror reflection of the RAMP for the leading DNA strand. In one Monte Carlo Step (MCS) each nucleotide of the gene sequence was drawn with a probability $p_{mut} = 0.01$, then substituted by another nucleotide with the probability described by the corresponding parameter in the substitution matrix. If the gene from the leading strand is under RAMP characteristic for it that means that the sense strand of the gene is under the RAMP specific for the leading DNA strand. If such a gene is inverted, it means that its sense strand is under the RAMP characteristic for the lagging DNA strand. After each round of mutations, we translated the nucleotide sequences into the amino acid sequences and compared the resulting composition of the proteins with the original. For each gene we calculated the selection parameter (T) for the amino acid composition which is the sum of absolute values of differences between fractions of amino acids as follows:

$$T = \sum_{i=1}^{20} |f_i(0) - f_i(t)|, \tag{1}$$

where: $f_i(0)$ is a fraction of a given amino-acid in the original sequence (before mutations) and $f_i(t)$ is a fraction of a given amino acid in the sequence after mutations in t MCS.

It describes the difference in the global amino acid composition of a protein coded by a given gene after mutations and its original sequence from the real genome. If T was below the assumed threshold, a gene stayed mutated and went to the next round of mutations (the next MC step). If T trespassed the threshold - the gene was "killed" and replaced by its allele from the second genomic sequence, originally identical, simulated parallely. As a value of the threshold we have assumed the average value T between 442 pairs of orthologs belonging to two related genomes: *B. burgdorferi* and *Treponema pallidum* which equals 0.3. These orthologs were extracted from the COGs database downloaded from *ftp://ftp.ncbi.nih.gov/pub/COG*. COGs contain protein sequences which are supposed to have evolved from one ancestral protein and usually fulfill a similar or same function [18]. The number of accumulated nucleotide substitutions and the number of gene replacements (the number of killed genes) were counted after each MCS. All simulations were performed for 1000 Monte Carlo steps, repeated 100 times and averaged. For comparison, the numbers of killed genes from different sets were normalized by the number of genes in the given set. In the simulations we have applied both stable and changing replication associated mutational pressure (RAMP). Stable RAMP means that during the whole simulation genes were subjected only to one pressure characteristic for the leading or the lagging strand. In the simulations with changing RAMP genes were alternately under the RAMP characteristic for the leading or the lagging DNA strand, changing with different frequencies. These simulations were carried out in different conditions described by the two parameters: F - the fraction of MC steps during the whole simulation in which the genes were subjected to mutational pressure characteristic for the strand on which they are normally located in the genome, N - Number of switches of the RAMP from leading to lagging one or *vice versa*. In sum, we have analyzed 87 different conditions of RAMP changing (different combinations of values F and N).

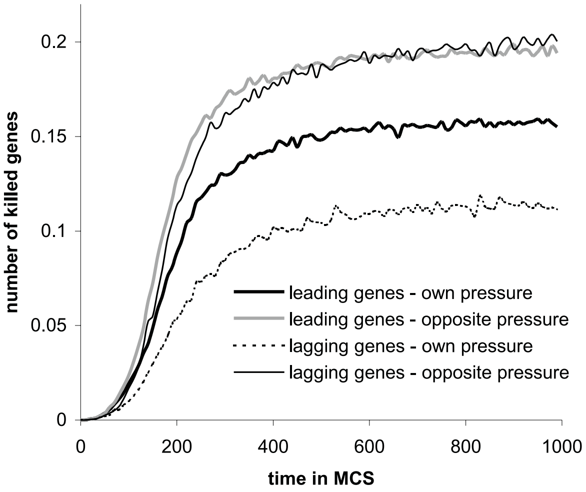

Fig. 2. The normalized number of killed genes from the leading and lagging strands of the *B. burgdorferi* genome. The genes were subjected to mutational pressure characteristic for them (their own pressure) and the mutational pressure characteristic for the complementary DNA strand (the opposite pressure). See text for details.

3 Results and discussion

Stable mutational pressure We have put all the genes from the leading and lagging DNA strands of the *B. burgdorferi* genome under the RAMP characteristic for the recent positions of these genes (Fig. 2), or under the mutational pressure characteristic for the complementary DNA strand. After simulations we found that:

- During simulations, the effect of killing grew in time and approximated to a relatively high level. This is because the sequences of genes tend to be better equilibrated with the mutational pressure but they are closer to the limit of tolerance so they are more often pushed outside this limit and killed.
- The killing effect for the genes staying under their own pressure is higher for the leading strand genes than for the lagging strand genes.
- Both sets of genes are better adapted to the mutational pressure characteristic for their recent positions in the genome than to the pressure from the opposite strand and the probability of killing them is lower at these positions. Furthermore, the killing effect under the opposite RAMP is equally deleterious for both sets of genes.

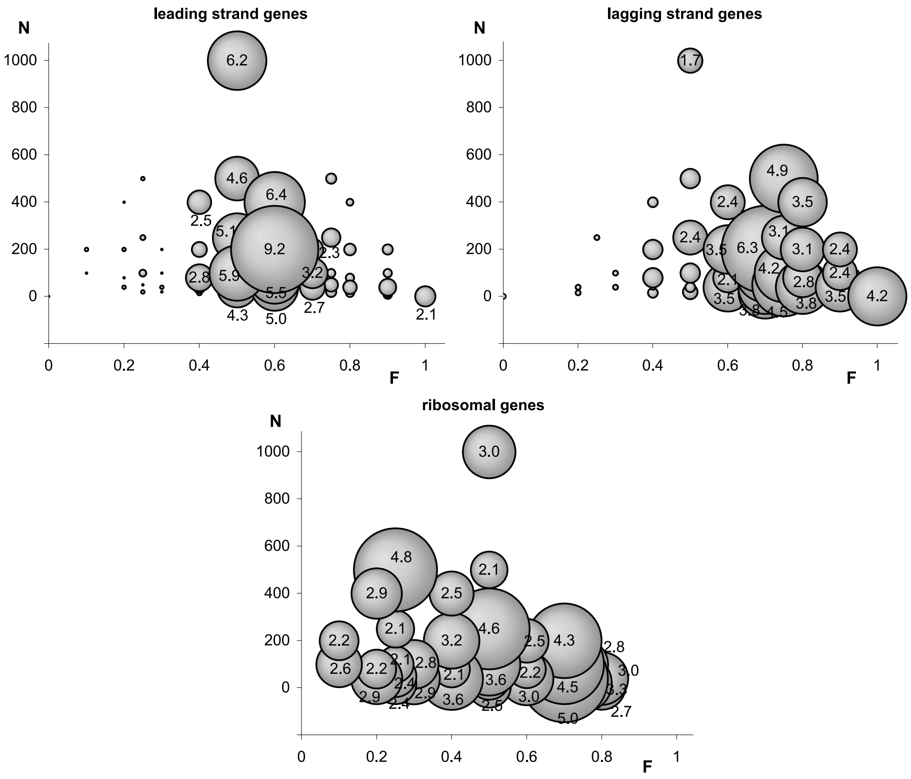

Fig. 3. Diagram presenting the best survival strategy for three sets of genes. This diagram shows which percent of a given set of genes has the highest survival chance under one of the 87 combinations of tested parameters (F and N) of changing mutational pressure after 1000 MCS of simulation. F - the fraction of MC steps during the whole simulation in which genes were subjected to mutational pressure characteristic for the strand on which they have been recently located in the genome; N - the number of switches of the mutational pressure from the leading to lagging one or *vice versa*.

Switching the mutational pressure In the earlier studies we have found that frequent changes of RAMP could be the best general strategy for gene survival [19]. This observation is consistent with the genetic observations that many genes or blocks of genes are organized in transposons which often change their positions in the bacterial chromosomes [20]. In the present studies we are showing the relationship between the frequency of gene transpositions (inversions) between differently replicating DNA strands and their survival. In Fig. 3 we have presented a diagram showing the best strategy for three sets of genes. This diagram shows which percent of a given set of genes has the highest survival chance under one of the 87 combinations of tested parameters (F and N) after 1000 MCS of simulation. Generally, genes prefer to stay longer under the RAMP to which they are actually subjected, but there are no preferred positions for the ribosomal genes located in the *B. burgdorferi* genome on the leading strand (almost all ribosomal genes in the overwhelming number of genomes are located on the leading DNA strands - [7]).

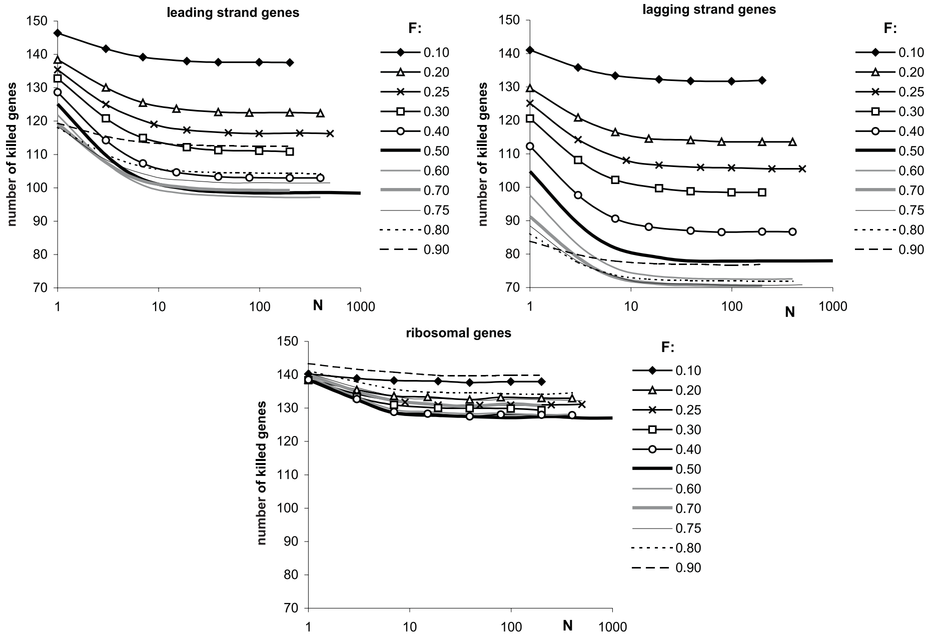

Fig. 4. Relationship between the number of killed genes and N (the frequency of switching the mutational pressure) for different F values (the fraction of MC steps in the whole simulation in which genes were subjected to the RAMP characteristic for the strand on which they have been recently located in the genome) for three sets of genes after 1000 MCS of simulation.

In Fig. 4 we have presented how the number of killed genes depends on N (the frequency of switching the mutational pressure) for different F values (the fraction of MC steps during the whole simulation in which genes were subjected to the RAMP characteristic for the strand on which they have been recently located in the genome). These analyses show that too frequent switching the direction of mutational pressure does not enhance significantly the gene survival. Usually switching every several hundreds of steps is close to the optimal gene survival.

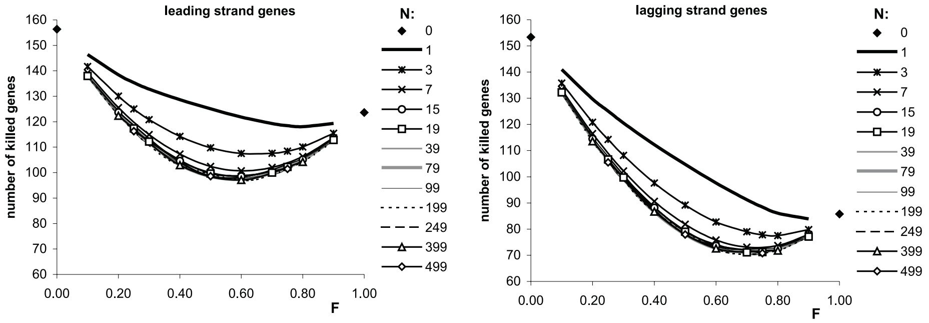

Fig. 5. Relationship between the number of killed genes and F (the fraction of MC steps in the whole simulation in which genes were subjected to the RAMP characteristic for the strand on which they have been recently located in the genome) for different N (the frequency of switching the mutational pressure) values for two sets of genes after 1000 MCS of simulation.

Relationship between the number of killed genes and F has a distinct minimum (Fig. 5). We have found that the killing effect for the leading strand genes is the lowest for F=0.59 and for the lagging strand genes for F=0.71 . Ribosomal genes do not profit as much from switching their positions. We have estimated that the frequency of gene elimination corresponds to about one per 10^9 replication cycles in Nature. This finding seems to support the genetic observations that gene translocation in prokaryotic genomes is under a rather strict control and many genes in closely related strains stay at the same positions ([21], [22]). It is obvious that staying under the same directional mutational pressure shifts the DNA nucleotide composition closer to the equilibrium state. The mutation rate of the DNA in equilibrium with the mutational pressure is lower than that of the DNA far from the equilibrium. That is why we observe another interesting effect of the switching mutational pressure from time to time. The number of mutations which are introduced into DNA sequences after switching the mutational pressure is higher than in the simulations where the mutational pressure is stable.

As it can be seen in Fig. 6 the number of accepted amino acid substitutions in coded proteins per site (substitutions which did not eliminate the gene function) is also higher. That means that the observed divergence of genes which recently changed their positions on chromosome should be higher, which was actually observed in numerous genomic analyses ([23] – [26]). In Fig. 6 it is also clear that the number of accepted substitutions is the lowest for the ribosomal proteins which are actually extremely conserved. The last observations, these from simulations as well as from genome analyses lead to the conclusion that switching the direction of the mutational pressure does not diminish the total frequency of mutations but rather introduces intragenic suppression mutations which complement the former mutations in the same gene. Such intragenic suppression should be much more effective for longer genes (see accompanying paper).

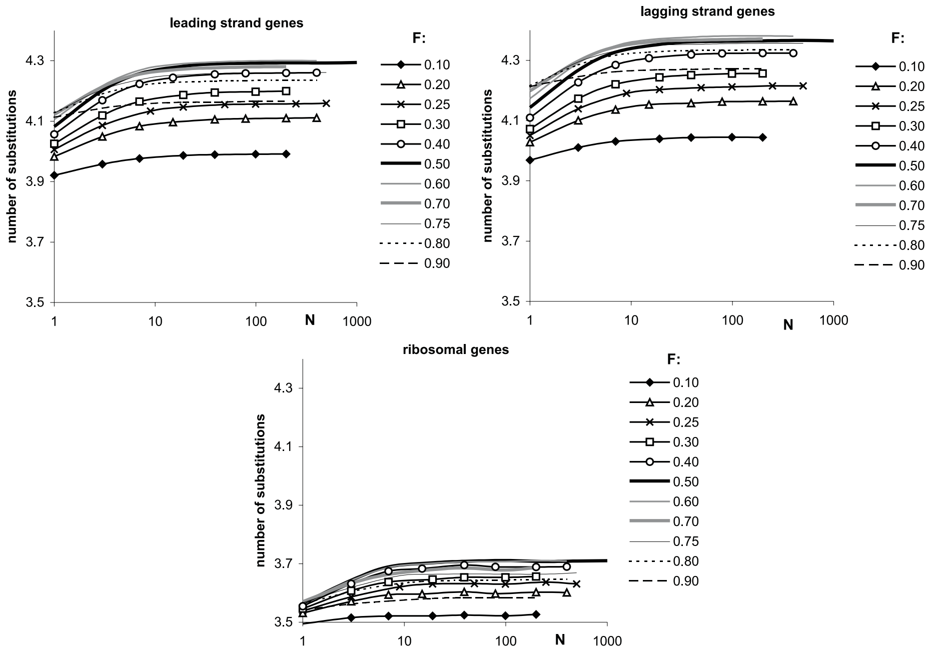

Fig. 6. Relationship between the number of accepted amino acid substitutions in coded proteins per site and N (the frequency of switching the mutational pressure) for different F values (the fraction of MC steps in the whole simulation in which genes were subjected to RAMP characteristic for strand on which they have been recently located in the genome) for three sets of genes after 1000 MCS of simulation.

The behavior of ribosomal genes needs special attention. These genes, in all the genomes analyzed thus far are usually located on the leading strand [7]. Our simulations have shown that they do not profit very much from transpositions (switching the mutational pressure) and the deleterious effect of the prolonged opposite mutational pressure is the same for the leading and lagging DNA strands. Thus, why do ribosomal genes prefer to stay at the leading strand? The answer could be in the topology of the transcription and replication. Since these genes are very intensively transcribed it could be important for them to concert the direction of replication fork movement and the direction of transcription. This could eliminate the possible deleterious effect of head on collisions of replication and transcription complexes ([27], [28]). The location of sense strands of these genes on the leading strand eliminates this effect.

Acknowledgemetns
The work was supported by the grant number 1016/S/IMi/03 and is done in the frame of COST Action P10 program. M.K. was supported by the Foundation for Polish Science.

References

1. Chargaff, E.: Chemical specificity of nucleic acids and mechanism of their enzymatic degradation. Experientia **6** (1950) 201–240
2. Lobry, J.R.: Properties of a general model of DNA evolution under no-strand-bias conditions. J. Mol. Evol. **40** (1995) 326–330, **41** 680
3. Frank, A.C., Lobry, J.R.: Asymmetric substitution patterns: a review of possible underlying mutational or selective mechanisms. Gene **238** (1999) 65–77
4. Kowalczuk, M., Mackiewicz, P., Mackiewicz, D., Nowicka, A., Dudkiewicz, A., Dudek, M.R., Cebrat, S.: DNA asymmetry and the replicational mutational pressure. J. Appl. Genet. **42** (2001) 553–577
5. Lobry, J.R.: Asymmetric substitution patterns in the two DNA strands of bacteria. Mol. Biol. Evol. **13** (1996) 660–665
6. Freeman, J.M., Plasterer, T.N., Smith, T.F., Mohr SC: Patterns of genome organization in bacteria. Science **279** (1998) 1827
7. McLean, M.J., Wolfe, K.H., Devine, K.M.: Base composition skews, replication orientation, and gene orientation in 12 prokaryote genomes. J. Mol. Evol. **47** (1998) 691–696
8. Mrazek, J., Karlin, S.: Strand compositional asymmetry in bacterial and large viral genomes. Proc. Natl. Acad. Sci. USA **95** (1998) 3720–3725
9. Mackiewicz, P., Gierlik, A., Kowalczuk, M., Dudek, M.R., Cebrat, S.: How does replication-associated mutational pressure influence amino acid composition of proteins? Genome Res. **9** (1999) 409–416
10. Rocha, E.P., Danchin, A., Viari, A. Universal replication biases in bacteria. Mol. Microbiol. **32** (1999) 11–16
11. Tillier, E.R., Collins, R.A. The contributions of replication orientation, gene direction, and signal sequences to base-composition asymmetries in bacterial genomes. J. Mol. Evol. **50** (2000) 249257
12. Mackiewicz, P., Kowalczuk, M., Mackiewicz, D., Nowicka, A., Dudkiewicz, M., Łaszkiewicz, A., Dudek, M.R., Cebrat, S.: Replication associated mutational pressure generating long-range correlation in DNA. Physica A **314** (2002) 646–654
13. Shepherd, J.C.: Method to determine the reading frame of a protein from the purine/pyrimidine genome sequence and its possible evolutionary justification. Proc. Natl. Acad. Sci. USA **78** (1981) 1596–1600
14. Smithies, O., Engels, W.R., Devereux, J.R., Slightom, J.L., Shen, S.H.: Base substitutions, length differences, and DNA strand asymmetries in the human Gg and Ag fetal globin gene region. Cell **26** (1981) 345–353
15. Cebrat, S., Dudek, M.R., Mackiewicz, P., Kowalczuk, M., Fita, M.: Asymmetry of coding versus non-coding strands in coding sequences of different genomes. Microb. Comp. Genomics **2** (1997) 259–268

16. 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.: Genomic sequence of a Lyme disease spirochaete, *Borrelia burgdorferi.* Nature **390** (1997) 580–586
17. Kowalczuk, M., Mackiewicz, P., Mackiewicz, D., Nowicka, A., Dudkiewicz, M., Dudek, M.R., Cebrat, S.: High correlation between the turnover of nucleotides under mutational pressure and the DNA composition. BMC Evol. Biol. **1** (2001) (1):13
18. 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.: The COG database: new developments in phylogenetic classification of proteins from complete genomes. Nucleic Acids Res. **29** (2001) 22–28
19. Dudkiewicz, M., Mackiewicz, P., Nowicka, A., Kowalczuk, M., Mackiewicz, D., Polak, N., Smolarczyk, K., Dudek, M.R., Cebrat, S.: Properties of Genetic Code under Directional, Asymmetric Mutational Pressure. In: Sloot PMA et al. (Eds.): Computational Conference ICCS 2003, Melbourne and St. Petersburg, June 2-4, 2003, LNCS **2657** (2003) 343–350
20. Chandler, M., Galas, D.J.: Cointegrate formation mediated by Tn9 II: Activity of IS1 is modulated by external DNA sequences. J. Mol. Biol. **170** (1983) 61-91
21. Eisen, J.A., Heidelberg, J.F., White, O., Salzberg, S.L.: Evidence for symmetric chromosomal inversions around the replication origin in bacteria. Genome Biol. **1** (2000):research0011
22. Suyama, M., Bork, P.: Evolution of prokaryotic gene order: genome rearrangements in closely related species. Trends Genet. **17** (2001) 10–13
23. Tillier, E.R., Collins, R.A.: Replication orientation affects the rate and direction of bacterial gene evolution. J. Mol. Evol. **51** (2000) 459–463
24. Rocha, E.P., Danchin, A.: Ongoing evolution of strand composition in bacterial genomes. Mol. Biol. Evol. **18** (2001) 1789–1799
25. Szczepanik, D., Mackiewicz, P., Kowalczuk, M., Gierlik, A., Nowicka, A., Dudek, M.R., Cebrat, S.: Evolution rates of genes on leading and lagging DNA strands. J. Mol. Evol. **52** (2001) 426–433
26. Mackiewicz, P., Mackiewicz, D., Kowalczuk, M., Dudkiewicz, M., Dudek, M.R., Cebrat, S.: High divergence rate of sequences located on different DNA strands in closely related bacterial genomes. J. Appl. Genet. **44** (2003) 561– 584
27. Brewer, B.J.: When polymerases collide: replication and the transcriptional organization of the *E. coli* chromosome. Cell **53** (1988) 679–686
28. French, S.: Consequences of replication fork movement through transcription units *in vivo.* Science **258** (1992) 1362–1365