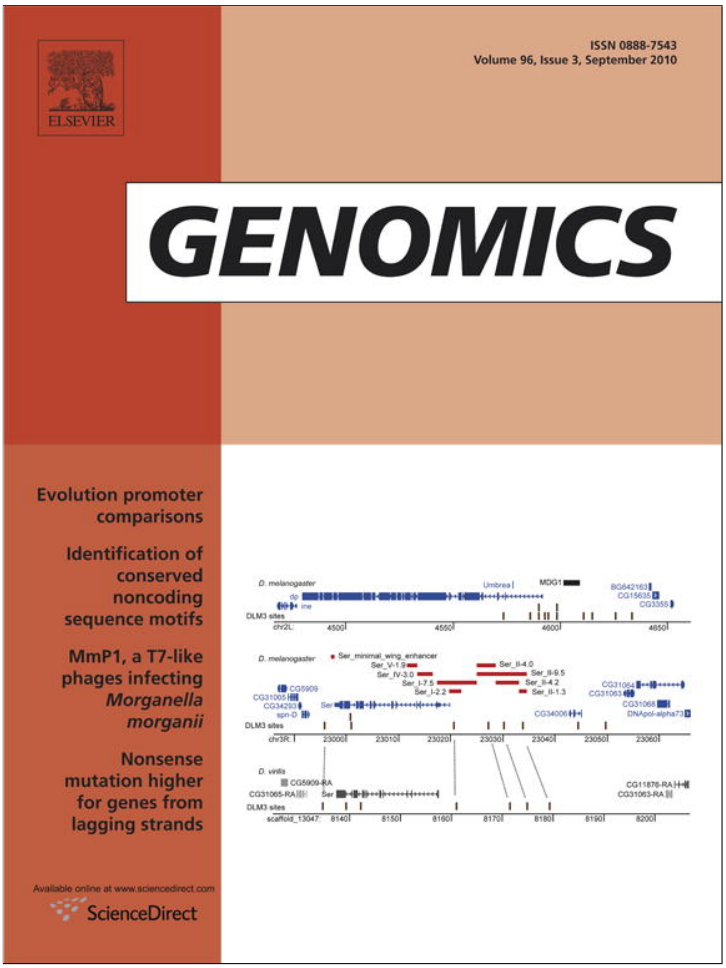

Genomics 96 (2010) 173–180

Contents lists available at ScienceDirect

Genomics

journal homepage: www.elsevier.com/locate/ygeno

The probability of nonsense mutation caused by replication-associated mutational pressure is much higher for bacterial genes from lagging than from leading strands

Vladislav Victorovich Khrustalev *, Eugene Victorovich Barkovsky

Department of General Chemistry, Belarussian State Medical University, Minsk, Dzerzinskogo, 83, Belarus

ARTICLE INFO

Article history:
Received 9 February 2010
Accepted 12 June 2010
Available online 18 June 2010

Keywords:
Replichore
Chirochore
Replication-associated mutational pressure
Preterminal codon usage
8-oxo-G

ABSTRACT

We studied nucleotide usage biases in 4-fold degenerated sites of all the genes from leading and lagging strands of 30 bacterial genomes. The level of guanine in 4-fold degenerated sites (G4f) is significantly lower in genes from lagging strands than in genes from leading strands, probably because of the faster rates of guanine oxidation in single-stranded DNA leading to G to T transversions. The rates of cytosine deamination causing C to T transitions are also higher in lagging strands. We showed that the level of codons able to form stop-codons by the way of G to T transversions and C to T transitions is always higher than the level of codons able to form stop-codons by the way of C to A transversions and G to A transitions. This circumstance can be an explanation of the lower percent of ORFs in lagging strands of bacterial replichores than in leading strands.

Introduction

There are two facts known about bacterial replichores (chirochores) since their discovery [1]. The first fact is the following. Nucleotide content of the leading strands of bacterial DNA is different from the nucleotide content of the lagging strands [1–3]. There is a single origin of replication (OriC) in bacterial "chromosome" [4] and a single region of its termination (ter) [5]. In fact, the same DNA strand is leading downstream the OriC and it is lagging upstream the OriC. So, bacterial genomic DNA is separated into two parts of an opposite "chirality" (in terms of nucleotide content) by "OriC" and "ter" regions [1,2].

In this work we confirmed that there is a difference in nucleotide content of leading and lagging strands [1–3,6,7] and reported that some general features of the nucleotide content distribution between leading and lagging strands of bacteria do exist. We concentrated our attention on the differences in nucleotide usage in 4-fold degenerated sites, while classical GC-skews analysis is based on calculation of total nucleotide content [1–3,7]. All the possible nucleotide mutations are synonymous in 4-fold degenerated sites. So, the level of cytosine in 4-fold degenerated sites (C4f) and the level of guanine in them (G4f) are the most sensitive indicators of replication-associated mutational pressure [6].

The second fact known about bacterial replichores is that the density of open reading frames is higher for the leading strand than for the lagging strand of each replichore [2,6–8]. In other words, for every "replichore" the percent of coding regions situated on leading strand is higher than the percent of coding regions situated on lagging strand. In our opinion, the same mutational processes should be responsible for nucleotide usage bias observed between leading and lagging strands and for the difference in the density of coding regions between leading and lagging strands.

How can the open reading frame disappear? It can disappear by the way of nonsense mutation. What is the substrate for nonsense mutations? There are so-called preterminal codons which can become terminal by the way of a single nucleotide substitution [9,10]. The level of preterminal codon usage shows the inverse linear correlation with G+C of bacterial genes [10]. There are codons which can become terminal by the way of C to T transition and those which can become terminal by the way of G to A transition. C to T transitions (caused by cytosine deamination) should be more frequent in lagging strands [11]. It means that G to A transitions should be more frequent in leading strands.

In this work we have shown that the usage of codons which can become terminal by the way of C to T transition (PCU C to T) is always higher than the usage of codons which can become terminal by the way of G to A transition (PCU G to A). So, the probability of nonsense transition is higher for ORFs from lagging strands, than for ORFs from leading strands.

Similar situation has been discovered by us for nonsense transversions. The usage of codons which can become terminal by the way of G to T transversion (PCU G to T) is always higher than the usage of codons which can become terminal by the way of C to A transversion (PCU C to A). G to T transversions caused by the oxidation of guanine should be more frequent in ORFs from lagging strands than in ORFs from leading strands [12], unlike C to A transversions.

We came to the conclusion that the nature of genetic code and the predecessor's effect are responsible for the higher amount of the

* Corresponding author. Minsk, 220029, Communisticheskaya 7-24, Belarus.
E-mail address: vvkhrustalev@mail.ru (V.V. Khrustalev).

0888-7543/$ – see front matter © 2010 Elsevier Inc. All rights reserved.
doi:10.1016/j.ygeno.2010.06.002

substrate for nonsense C to T and G to T mutations (in comparison with the substrate for nonsense G to A and C to A mutations) in all coding districts. Biochemical causes of replication-associated mutational pressure (RAMP) are responsible for the higher frequencies of C to T and G to T mutations in genes from lagging strands.

Results

Nucleotide usage biases in 4-fold degenerated sites of genes from lagging and leading strands

We studied 30 completely sequenced bacterial genomes with different levels of total GC-content. As one can see in Fig. 1, bacterial species from different phylums have been included in this work (Proteobacteria, Cyanobacteria, Firmicutes, Fusobacteria, Aquificae, Thermotogae, Spirochaetes, Deinococcus-Thermus).

Differences in nucleotide usage in 4-fold degenerated sites between genes from leading and genes from lagging strands for each replichore have been calculated (see Supplementary material, Tables 1–4). Although the difference in each nucleotide usage is not constant among all the replichores, general conclusions have been made after the appliance of paired differences test to the data obtained.

Level of G4f is significantly higher ($P<0.001$) in genes from leading strands than in genes from lagging strands; average difference is equal to $2.91 \pm 0.36\%$. Level of C4f is significantly lower ($P<0.001$) in genes from leading strands than in genes from lagging strands; average difference is equal to $3.54 \pm 0.37\%$.

The level of T4f is significantly higher ($P<0.05$) in genes from leading strands than in genes from lagging strands; however, an average difference is rather low; it is equal to $1.00 \pm 0.50\%$. There is no significant difference between the level of A4f in genes from leading and lagging strands.

The usage of codons able to form stop-codons by the way of a single nucleotide mutation occurred under the influence of replication-associated mutational pressure

The data described in previous subsection is consistent with experimental observations: the rates of cytosine and adenine deamination, as well as the rates of guanine and thymine oxidation

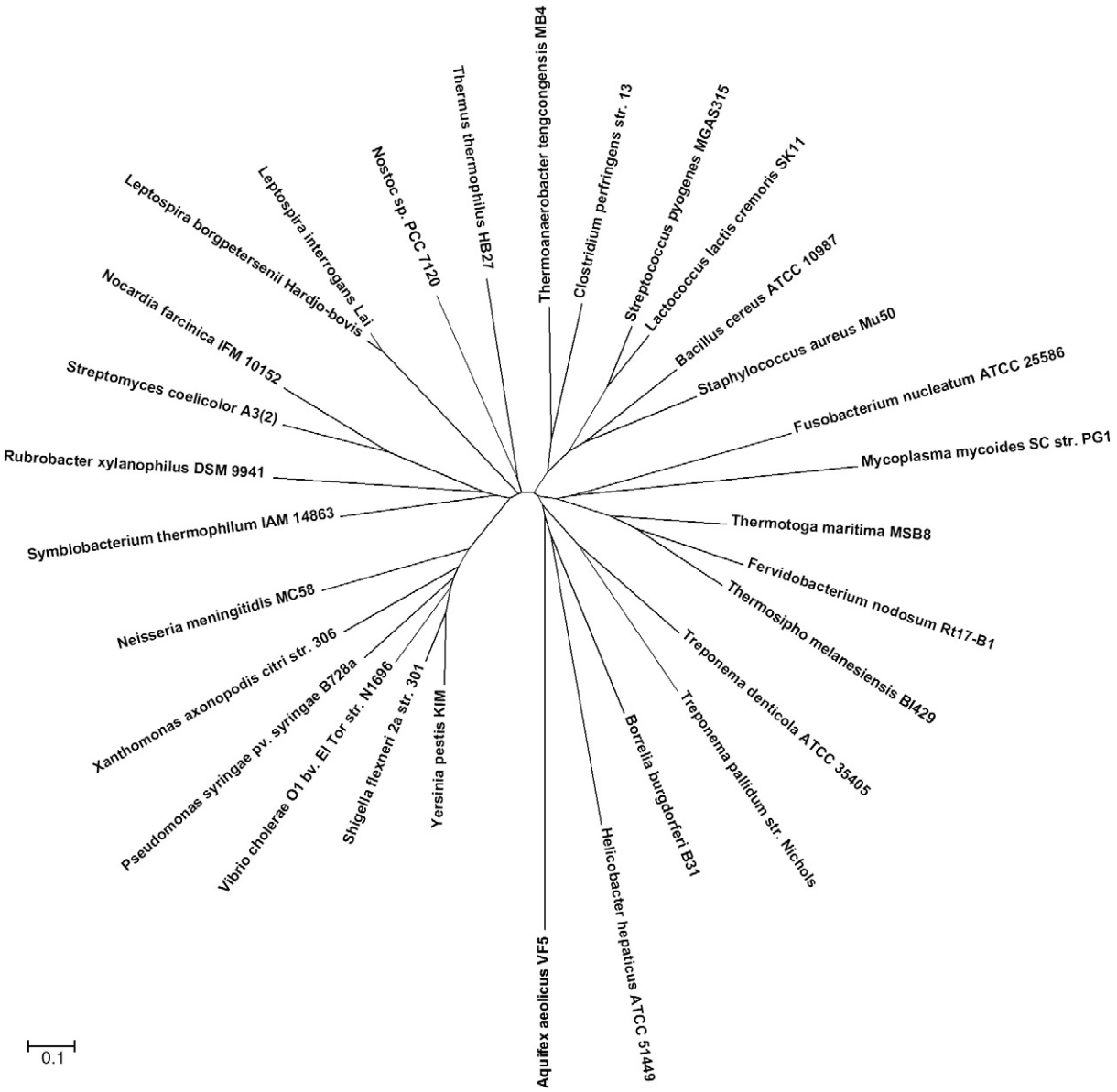

Fig. 1. NJ-tree built for amino acid sequences of DNA-polymerase I from bacterial species included in this study. PAM-matrix has been used for the alignment; evolutionary distances have been calculated by PAM-matrix (Dayhoff).

are higher in single-stranded DNA than in double-stranded DNA [11,12]. So, we can suggest four main molecular mechanisms of replication-associated mutational pressure.

The first mechanism is the deamination of cytosine preferably taking place in lagging strands [2] which finally causes C to T transitions in genes from lagging strands and G to A transitions in genes from leading strands. We counted the usage of codons able to form terminal codons by the way of C to T transition in genes from lagging strands (PCU C to T) and the usage of codons able to form terminal codons by the way of G to A transition in genes from leading strands (PCU G to A). As you can see in Fig. 2, the level of "PCU C to T" in genes from lagging strands is always much higher than the level of "PCU G to A" in genes from leading strands (average ratio is equal to 3.6 ± 0.2). It means that the probability of nonsense mutation caused by C to T transition that occurred in lagging strand is much higher for genes situated on lagging strand than for genes situated on complementary leading strand (C to T transition in lagging strand is inherited by the leading strand as G to A transition).

The second mechanism of replication-associated mutational pressure is the oxidation of guanine in lagging strands (finally causing G to T transversions in lagging strands and C to A transversions in leading strands). We counted the usage of codons able to form terminal codons by the way of G to T transversion in genes from lagging strands (PCU G to T) and the usage of codons able to form terminal codons by the way of C to A transversion in genes from leading strands (PCU C to A). As one can see in Fig. 3, the level of "PCU G to T" in genes from lagging strands is always much higher than the level of "PCU C to A" in genes from leading strands (average ratio is equal to 2.4 ± 0.1). It means that the probability of nonsense mutation caused by G to T transversion occurred in lagging strands is much higher for genes from lagging strands than the probability of nonsense mutation caused by C to A transversion for genes from complementary leading strands.

The third mechanism of replication-associated mutational pressure should be the oxidation of thymine more frequently tasking place in lagging strands. This mutation results in T to C transition in lagging strand and in A to G transition in leading strand [13,14]. There is no nonsense mutation that can occur due to T to C or A to G transition in the universal genetic code.

The fourth mechanism of replication-associated mutational pressure is the deamination of adenine causing A to G transitions in lagging strands [13,14]. As we have written above, there is no nonsense mutation that can occur due to T to C or A to G transition in the universal genetic code. However, there is a possibility of A to T transversion occurrence due to adenine deamination in lagging strand (see Hypothetical molecular mechanisms of replication-associated mutational pressure). So, we counted the usage of codons able to form terminal codons by the way of A to T transversion in genes from lagging strands (PCU A to T) and the usage of codons able to form terminal codons by the way of T to A transversion in genes from leading strands (PCU T to A). In general, level of "PCU A to T" usage in genes from lagging strands is some higher ($P<0.01$) than the level of "PCU T to A" usage in genes from leading strands (average ratio is equal to 1.5 ± 0.1).

The difference in density of coding regions between leading and lagging strands is usually higher for GC-poor genomes

In four from six bacterial species with G+C higher than 60% (see Supplementary material, Table 1) the difference in density of coding regions for leading and lagging strand is relatively small (approximately 55% of coding regions on leading strand versus approximately 45% of coding regions on lagging strand). In our opinion, the lower are the rates of GC to AT mutations, the lower are the rates of nonsense mutations caused by single nucleotide substitution [10]. Efficient mechanisms of the reparation of C to U and G to 8-oxo-G mutations should prevent the most of the nonsense mutations caused by them.

In GC-poor genomes (G+C<40%) of *Streptococcus pyogenes*, *Thermoanaerobacter tengcongensis*, *Lactococcus lactis*, *Bacillus cereus*, *Staphylococcus aureus* and *Clostridium perfringens* the percent of coding regions situated on lagging strands is more than 2.5 times lower than the percent of coding regions situated on leading strands (see Supplementary material, Tables 3 and 4). This feature can be caused by the high rates of C to U and G to 8-oxo-G mutations. Insufficient reparation of these mutations should be responsible for the decrease of total GC-content in genomes of these bacterial species [13]. As we have discussed before, both C to U and G to 8-oxo-G mutations occur preferably in lagging strands, so they should cause a lot of nonsense mutations preferably in ORFs from lagging strands.

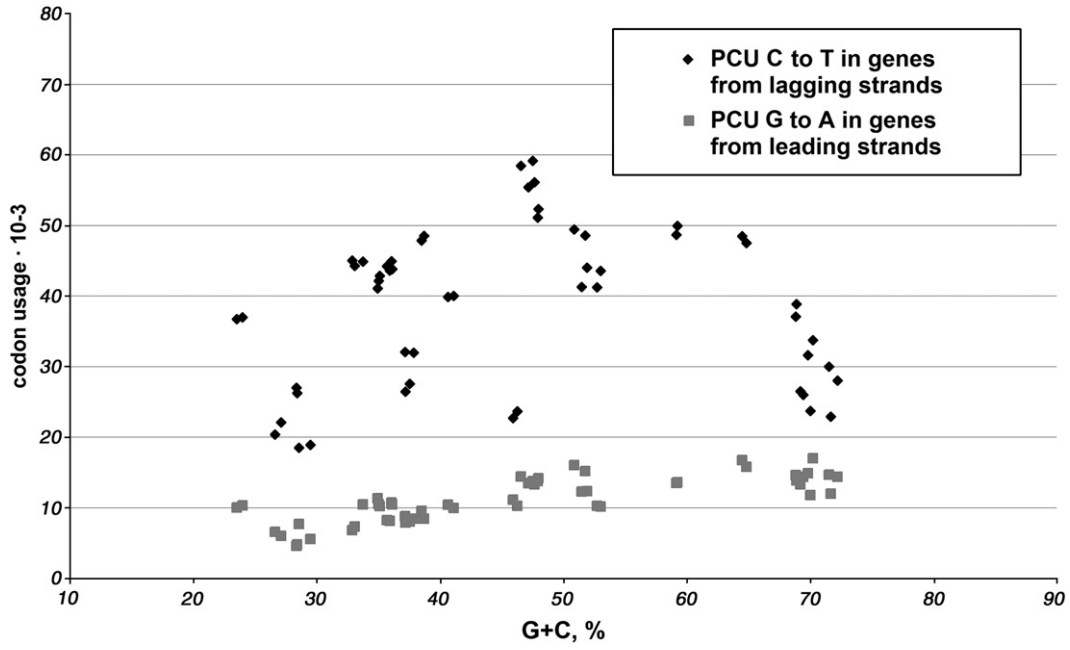

Fig. 2. Dependence of average levels of preterminal codons able to become stop-codons by the way of C to T transition (PCU C to T) in genes from lagging strands and average levels of preterminal codons able to become stop-codons by the way of G to A transition (PCU G to A) in genes from leading strands on the average level of G+C.

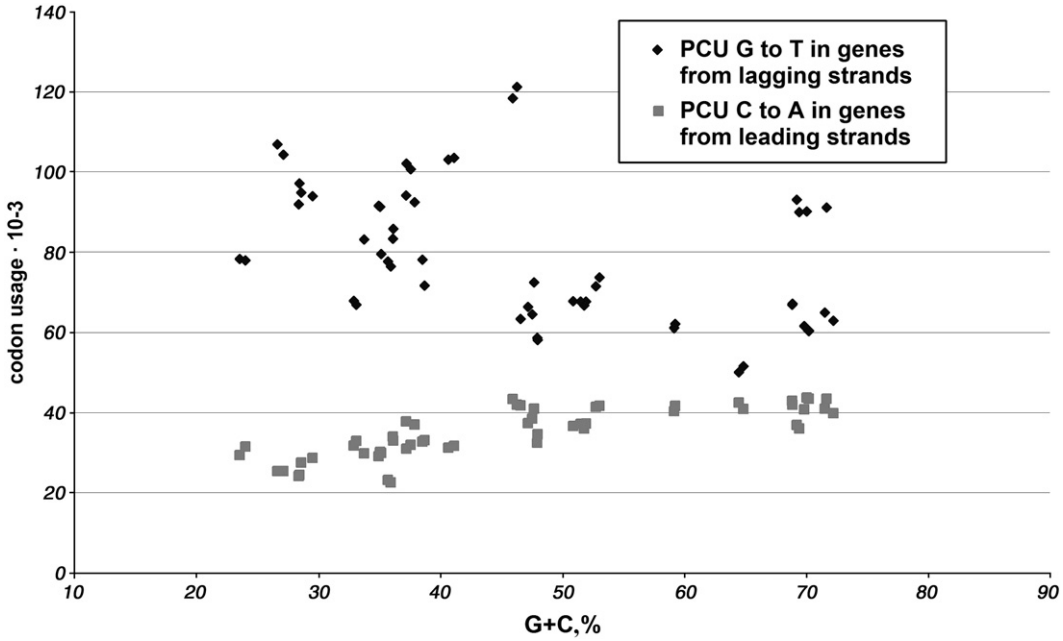

Fig. 3. Dependence of average levels of preterminal codons able to become stop-codons by the way of G to T transversion (PCU G to T) in genes from lagging strands and average levels of preterminal codons able to become stop-codons by the way of C to A transversion (PCU C to A) in genes from leading strands on the average level of G+C.

Inversions temporary disturb the structure of replichores

The level of G4f is higher in ORFs from leading strands than in ORFs from lagging strands in the most of chirochores. One of the chirochores of *Thermotoga maritima* (see Fig. 4) provides us with an interesting exception from this rule: there is no significant difference between G4f in ORFs from leading and lagging strand of this replichore.

Several inversions have been found by us in Watson strand of *T. maritima* (see Fig. 4). These inversions not only disturbed a bias in nucleotide usage, but also led to the identical density of coding regions on both leading and lagging strands from the given *T. maritima* replichore. The level of G4f is higher than the level of C4f only in genes from leading strand of the "normal" *T. maritima* replichore (see Fig. 4). There is no difference between G4f and C4f in genes from lagging strand of the "normal" *T. maritima* replichore. As you can see in Fig. 4,

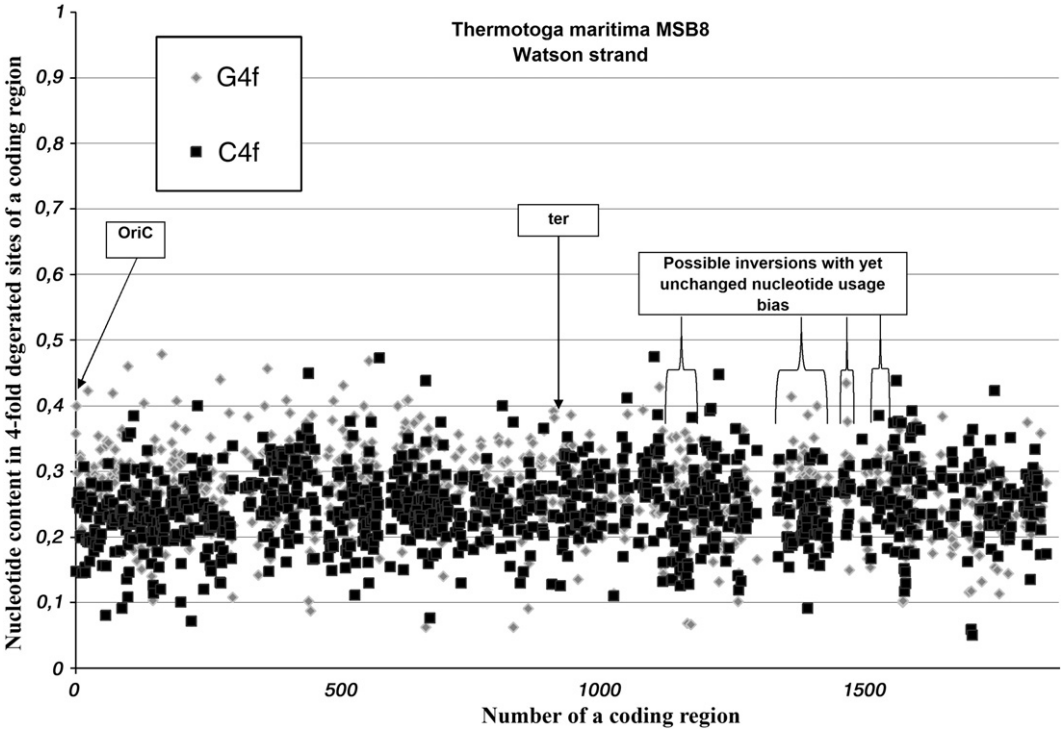

Fig. 4. Levels of guanine and cytosine content in 4-fold degenerated sites (G4f and C4f, respectively) in all coding regions along the length of Watson strand from *T. maritima* MSB8 genome.

there are at least four regions of the "mutated" *T. maritima* replichore with the bias (G4f > C4f) characteristic for the genes from complementary strand. This fact makes us hypothesize that these four regions with deviated bias are relatively large inversions.

The percent of ORFs is usually higher for leading strand, than for lagging strand [8]. However, we have found bacterial specie with the deviation form this rule. This specie is *T. thermophilus* (see Fig. 5). In one of its replichores the percent of coding regions situated on leading strand is equal to 47%, while the percent of coding regions situated on lagging strand is equal to 53%.

Large possible inversion has been found by us in the genome of *T. thermophilus* (see Fig. 5). This inversion is an interesting one because the bias in nucleotide usage inside it is close to that in other parts of the same replichore, while the density of coding regions is not. The deviation from the universal rule was caused only by this inversion: the density of coding regions is lower for the rest of the lagging strand than for the rest of the leading strand of this *T. thermophilus* replichore (see Fig. 5).

In general, this phenomenon can be interpreted with the help of Kimura's theory of neutral molecular evolution [15]. The inversion leads to the disturbance of both nucleotide usage bias and coding region density distribution between leading and lagging strands [16]. The process of replication-associated mutational pressure should result in the "improvement" of the local nucleotide usage bias disturbance. Indeed, all the mutations in 4-fold degenerated sites are neutral [15]. So, the speed of the nucleotide usage bias "improvement" should depend on the intensity of replication-associated mutational pressure and on the random genetic drift [15].

Nonsense mutation can be neutral, for example, in case if it happens inside the copy of functional gene occurred due to duplication [10,15]. There also can be genes coding for relatively unnecessary products [10]. In general, some of nonsense mutations may be fixed by random genetic drift [10,15]. However, the most of nonsense mutations should bring negative consequences [15]. The speed of the coding region density "improvement" should be slower than the speed of the nucleotide usage bias "improvement". Anyway, sooner or later, the density of coding region distribution between leading and lagging strands becomes "normal". In our opinion, duplications should play significant role in this process. Additional copies of genes should persist in leading strands longer than in lagging strands, in which they are at a greater risk of nonsense mutation.

Consequences of OriC and ter translocations

In chromosomes of *Thermosipho melanesiensis* and *Fervidobacterium nodosum* location of OriC provided by DoriC database [4] is different from that determined by GC-skews analysis. In our opinion, OriC region has changed its position in these two chromosomes in a relatively recent past. That is why both nucleotide usage biases and density of coding regions have not already been changed. As one can see in Fig. 6, the most of the "present" leading strand of *T. melanesiensis* is guanine-poor and the percent of coding regions (41%) is low for it. Interestingly, present position of OriC in *T. melanesiensis* chromosome is close to the previous position of ter region. Position of ter should have been changed too. If positions of OriC and ter in *T. melanesiensis* chromosome will not be changed for a long period of time, guanine usage in 4-fold degenerated sites for genes from lagging and genes from leading strands will become practically equal to each other. After the period of "temporary equality", biases will become "normal."

Probably, the absence of nucleotide usage biases in replichores of *Nostoc sp.* [3] and *Aquifex aeolicus* is the consequence of OriC and ter translocation similar to that in *T. melanesiensis* (see Fig. 6), but occurred in a relatively distant past. So, nucleotide usage biases in genes from lagging and leading strands of *Nostoc sp.* and *Aquifex aeolicus* seem to become approximately equal to each other. Probably, after a certain amount of generations *Nostoc sp.* and *Aquifex aeolicus* will acquire "normal" nucleotide usage biases in their replichores.

Genome of *Streptomyces coelicolor* also demonstrates reversed nucleotide usage bias (G4f is some higher and C4f is some lower for genes from lagging strands), while the density of coding regions is normal for its leading and lagging strands (see Supplementary

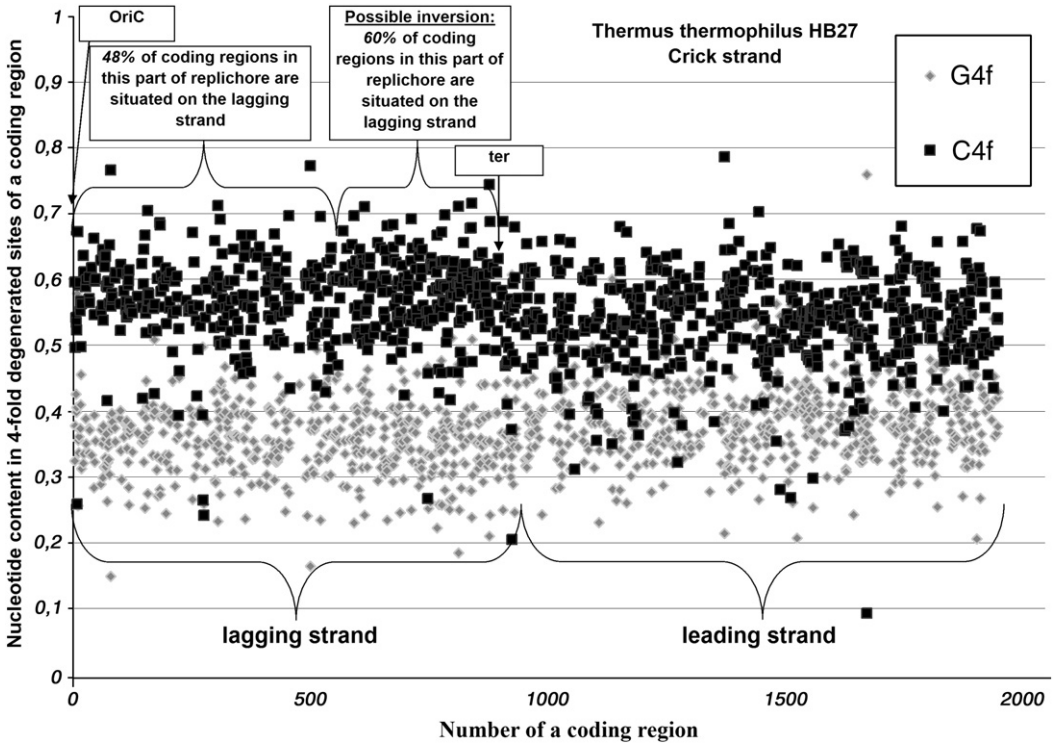

Fig. 5. Levels of guanine and cytosine content in 4-fold degenerated sites (G4f and C4f, respectively) in all coding regions along the length of Crick strand from *T. thermophilus* HB27 genome.

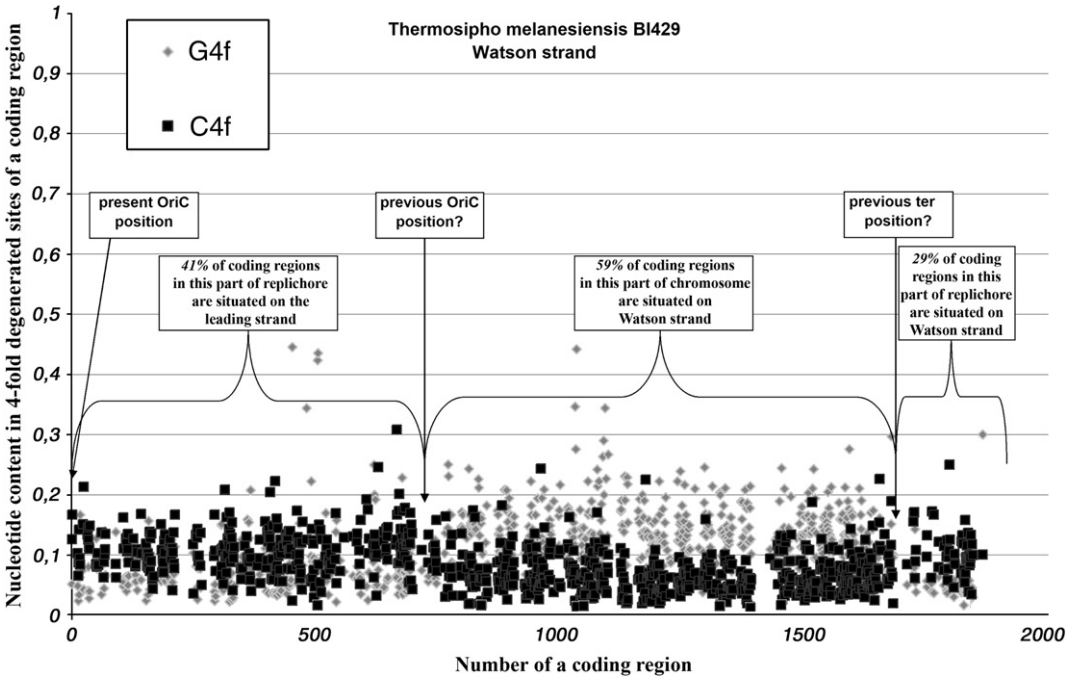

Fig. 6. Levels of guanine and cytosine content in 4-fold degenerated sites (G4f and C4f, respectively) in all coding regions along the length of Watson strand from *Thermosipho melanesiensis* BI429 genome.

material, Table 1). Normally, the chromosome of *Streptomyces coelicolor* is linear, but its circularization accompanied by large deletions and amplifications occurs frequently [17]. In our opinion, frequent chromosome rearrangements described for *Streptomyces* species and the absence of ter region should increase the probability of OriC translocation.

Discussion

Hypothetical molecular mechanisms of replication-associated mutational pressure

Our data showed that there is a process decreasing the level of guanine in lagging strands of bacterial replichores. It was shown that the rates of guanine oxidation are higher in single-stranded DNA, than in double-stranded DNA [12]. Lagging strand exists in a single-stranded form during the replication for a longer period of time than leading strand [11]. The rates of guanine oxidation should be higher in lagging strand than in leading strand. In our opinion, the level of G4f in genes from lagging strands is usually lower than in genes from leading strands because of the higher rates of guanine oxidation in lagging strands.

During the replication adenine is frequently incorporated in front of oxidized guanine (8-oxo-G) [13]. Normally, adenine should be excised from newly synthesized DNA duplex by MutY reparative enzyme which recognizes 8-oxo-G:A mispairs [13]. Then cytosine should occur in front of 8-oxo-G and the later one should be replaced by guanine [13]. This reparative pathway is antimutagenic in case of guanine oxidation in lagging strands, while it is promutagenic in case of 8-oxo-G incorporation into the growing strand of DNA [13]. So, if MutY expression and function is not altered in a given bacteria, its GC-content should grow due to AT to GC transversions and the difference in nucleotide usage and coding region density between leading and lagging strands should be lower. However, the higher is the probability of guanine oxidation in lagging strands, the higher is the probability that A:8-oxo-G mispair will not be repaired by MutY till the next round of replication.

Moreover, 8-oxo-G:A mispair may also be recognized and repaired by MutM enzyme which excises 8-oxo-G [13]. However, normally, MutM preferably excises 8-oxo-G from 8-oxo-G:C and 8-oxo-G:G mispairs [13].

According to the experimental works, the rates of cytosine deamination should be higher for single-stranded DNA than for double-stranded DNA [11]. Cytosine deamination in lagging strands seems to be inescapable. Deaminated cytosine (uracil) may be excised from single-stranded DNA by Uracil-DNA glycosilases [13]. However, DNA-polymerases usually incorporate adenine in front of AP-sites (according to "A-rule"). In this case, thymine occurs in place of excised uracil [13].

The level of C4f is usually higher in genes from lagging strands. Probably, there are mutational processes increasing the level of cytosine in lagging strands and "hiding" the effect of frequent cytosine deamination. Thymine can be oxidized in DNA forming 5-formyl-uracil which frequently mispairs with guanine [13,14]. The rates of thymine oxidation are also higher in single-stranded DNA than in double-stranded DNA [12]. So, the level of cytosine in lagging strands may be increased due to the frequent T to C mutations occurrence caused by thymine oxidation.

Adenine deamination in single-stranded DNA should lead not only to A to G but also to A to T mutations. Reparative enzyme AlkA is able to excise deaminated adenine (hypoxanthine) from single-stranded DNA [14]. If DNA-polymerase will incorporate adenine in front of AP-site left after hypoxanthine excision, A to T transversion will occur.

One should remember that DNA-polymerase I itself (see Fig. 1) is also involved in base excision repair [18]. So, features of this enzyme should modulate replication-associated mutational pressure.

Biases in different preterminal codons usage and universal genetic code

In this work we reported that some features of codon usage distribution cannot be altered by directed mutational pressure. Namely, the level of codons able to form terminal codons by the way of C to T transition (PCU C to T) is always higher than the level of codons able to form terminal codons by the way of G to A transition

(PCU G to A) in bacterial genomes either with G+C equal to 23.8% or in those with G+C equal to 72.3%.

"PCU C to T" are codons coding for glutamine (CAA and CAG) and one from the six codons coding for arginine (CGA). "PCU G to A" is a single codon coding for tryptophan (TGG). In mycoplasma species tryptophan is encoded by two codons (TGG and TGA) [10]. The last one (which is suppressed stop-codon) is also able to form terminal codon by the way of G to A transition. The level of tryptophan usage is always low in proteins from all the living species and it is usually lower than the level of glutamine [10].

Codons able to form terminal codons by the way of G to T transversion (PCU G to T) are those coding for glutamic acid (GAA and GAG) and one from the four codons coding for glycine (GGA). Codons able to form terminal codons by the way of C to A transversion (PCU C to A) are two from the six codons coding for serine (TCA and TCG), one from the two codons coding for tyrosine (TAC) and one from the two codons coding for cysteine (TGC). It seems like the level of aspartic acid usage is usually high in bacterial proteins, while levels of tyrosine and, especially, cysteine are low [10].

So, the nature of genetic code and the common predecessor's effect are responsible for biases in different preterminal codons usage.

Materials and methods

In this study we used 30 bacterial genomes as the material. To study them we chose to analyze the information on codon usage in each coding region of the genome stored in Codon Usage Database (www.kazusa.or.jp/codon) [19]. Bacterial species and names of their reference strains are written below: *Streptomyces coelicolor* A3 (2); *Rubrobacter xylanophilus* DSM 9941; *Nocardia farcinica* IFM 10152; *T. thermophilus* HB27; *Symbiobacterium thermophilum* IAM 14863; *Xanthomonas axonopodis* pv. citri str. 306; *Pseudomonas syringae* pv. syringae B728a; *Treponema pallidum* subsp. pallidum str. Nichols; *Neisseria meningitidis* MC58; *Shigella flexneri* 2a str. 301; *Yersinia pestis* KIM; *Vibrio cholerae* O1 biovar eltor str. N16961; *T. maritima* MSB8; *Leptospira borgpetersenii* serovar Hardjo-bovis L550; *S. pyogenes* MGAS315; *T. tengcongensis* MB4; *Treponema denticola* ATCC 35405; *Lactococcus lactis* subsp. cremoris SK11; *Helicobacter hepaticus* ATCC 51449; *Bacillus cereus* ATCC 10987; *Leptospira interrogans* serovar Lai str. 56601; *Staphylococcus aureus* subsp. aureus Mu50; *Clostridium perfringens* str. 13; *Borrelia burgdorferi* B31; *Fusobacterium nucleatum* subsp. nucleatum ATCC 25586; *Mycoplasma mycoides* subsp. mycoides SC str. PG1; *Aquifex aeolicus* VF5; *Nostoc sp.* PCC 7120; *T. melanesiensis* BI429; *F. nodosum* Rt17-B1.

Our aim was to study bacterial species from different phylogenetic clades with different average GC-content in their genomes. Indeed, we included in this study species with extremely low (such as *Mycoplasma mycoides* with G+C = 23.8%) and extremely high GC-content (such as *Streptomyces coelicolor* with G+C = 72.3%), as well as many species with average GC-content. Phylogenetical relationships between bacterial species included in this study (Fig. 1) have been represented by NJ-dendrogram made for amino acid sequences of their DNA-polymerase I [18] constructed by MEGA 4 [20] (PAM evolutionary distances were calculated and PAM matrix was used for the alignment).

Special algorithm for the visualization of bacterial replichores and for all the necessary calculations has been written by us. This algorithm is called "Replichore Viewer" (www.barkovsky.hotmail.ru). This is a new version of "CGS" algorithm [21] with the additional module that performs calculations in coding regions from Watson and Crick strands separately.

To work with MS Excel spreadsheet called "Replichore Viewer" one should insert all the information from the "List of codon usage for each CDS" into its list "All CDSs". The "List of codon usage for each CDS" can be downloaded from Codon Usage Database [19]. The location of OriC for all the bacterial genomes studied has already been determined or predicted [4]. With the help of "Replichore Viewer" we determined approximate location of the replication termination region for every chromosome.

"Replichore Viewer" calculates nucleotide content in every codon position, as well as in 4-fold degenerated sites for every ORF. Then this algorithm calculates average level of all those indexes for ORFs from leading and for ORFs from lagging strands of each of the two replichores. For the appropriate work of "Replichore Viewer" one should enter the information on codon usage (from the "List of codon usage for each CDS") beginning from the OriC region. Then researcher needs to determine the location of the "ter" region and insert the number of the last line containing ORF from replichore 1 into the formulae from "Replichore Viewer" list (the complete step-by-step instruction is written in the list of this MS Excel spreadsheet).

"Replichore Viewer" also calculates usage of preterminal codons able to form terminal codons by the way of C to U transition, G to A transition, G to T transversion, C to A transversion, A to T transversion and T to A transversion in every ORF from leading and from lagging strands of each of the two replichores.

There are two prokaryotic "chromosomes" in *V. cholerae* as well as in *L. borgpetersenii* and *Leptospira interrogans*. We analyzed two chromosomes of *V. cholerae* separately, while we excluded from our study "minichromosomes" of Leptospiras because of their relatively short length.

There are numerous genomic islands with lower GC-content in the genome of *Neisseria meningitidis*. To avoid the disturbance of the average levels of nucleotide content we cut away two longest genomic islands from this genome. In the chromosome II of *V. cholerae* there is a long GC-poor genomic island that was also cut away during our calculations.

We came to the conclusion that there are no GC-skews which can be used to study replichores in the genome of *Nostoc sp.* PCC 7120 (this fact has already been discovered by Nikolaou and Almirantis [3]) and in the genome of *Aquifex aeolicus* VF5.

The location of OriC provided by DoriC [4] is not consistent with GC-skews for *T. melanesiensis* and *F. nodosum* genomes. We excluded *T. melanesiensis* and *F. nodosum* genomes from the group of analyzed bacterial replichores.

Finally, we used average levels of nucleotide usage in 4-fold degenerated sites from 54 lagging and 54 leading strands (from 54 replichores of 27 bacterial chromosomes) to perform paired differences test between them. The difference between the average level of each nucleotide usage in 4-fold degenerated sites of genes from leading strand and genes from lagging strand has been calculated for every replichore studied. Then we applied t-test to the 54 differences between each nucleotide usage in 4-fold degenerated sites of genes from lagging and genes from leading strands. Average differences have been calculated, statistical significance of each difference has been estimated.

Diagrams representing dependences of preterminal codon usage (PCU) from G+C have been built. We calculated the usage of codons able to form terminal codons by the way of C to T transition (PCU C to T) in ORFs from leading strands and the usage of codons able to form terminal codons by the way of G to A transition (PCU G to A) in ORFs from lagging strands. The usage of codons able to form terminal codons by the way of G to T transversion (PCU G to T) has been calculated in ORFs from leading strands and the usage of codons able to form terminal codons by the way of C to A transversion (PCU C to A) has been calculated in ORFs from lagging strands. We also calculated the usage of codons able to form terminal codons by the way of T to A transversion (PCU T to A) in ORFs from leading strands and the usage of codons able to form terminal codons by the way of A to T transversion (PCU A to T) in ORFs from lagging strands.

The significance of the difference between "PCU C to T" and "PCU G to A", between "PCU G to T" and "PCU C to A" as well as between "PCU A to T" and "PCU T to A" has also been tested by paired differences test.

Nucleotide usage in 4-fold degenerated sites of coding regions has already been used for the study of replication-associated mutational pressure [6]. However, in that study, as well as in other studies [1–3,7], one half of each DNA strand has been compared with another half of the same strand. In the present study we compared two strands of DNA from each replichore. This method sometimes may be helpful. For example, in such species as *T. thermophilus* and *T. maritima* one replichore is quite normal, while another one is altered by inversions. Comparison of one half of DNA strand with another half in *T. thermophilus* and *T. maritima* genomes is not appropriate since it leads to the incorrect conclusion that both replichores are altered.

Conclusions

Hypothesis of bacterial replichores maintenance based on the effects of asymmetric mutational pressure and negative selection has been formulated by us in the final part of this study.

There are several molecular mechanisms of replication-associated mutational pressure. All of them should be based on the higher rates of certain nucleotide mutations in single-stranded DNA. Replication-associated mutational pressure results at least in C to T, G to T, and T to C mutations in genes from lagging strands and in G to A, C to A and A to G mutations in genes from leading strands.

The level of preterminal codons able to become terminal codons by the way of C to T and G to T mutation is always much higher than the level of preterminal codons able to become terminal codons by the way of G to A and C to A mutation in the most of bacterial genes. It means that the probability for the mutation caused by replication-associated mutational pressure to be nonsense is much higher for genes from lagging strands than for genes from leading strands. That is why the density of ORFs is usually lower in lagging strands.

Inversions as well as translocations of OriC and ter regions disturb the structure of replichores. However, this disturbance is temporary. At first, replication-associated mutational pressure "corrects" the nucleotide usage bias in the inverted DNA. Then the gene duplication events accompanied by the unequal probability of nonsense mutation for genes situated in lagging and leading strands "correct" the density of ORFs in the inverted region.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.ygeno.2010.06.002.

References

[1] J.R. Lobry, Asymmetric substitution patterns in the two DNA strands of bacteria, Mol. Biol. Evol. 13 (1996) 660–665.
[2] J.R. Lobry, N. Sueoka, Asymmetric directional mutation pressures in bacteria, Genome Biol. 3 (2002) 0058.
[3] C. Nikolaou, Y. Almirantis, A study on the correlation of nucleotide skews and the positioning of the origin of replication: different modes of replication in bacterial species, Nucleic Acids Res. 33 (2005) 6816–6822.
[4] F. Gao, C.T. Zhang, DoriC: a database of *oriC* regions in bacterial genomes, Bioinformatics 23 (2007) 1866–1867.
[5] J.Z. Dalgaard, T. Eydmann, M. Koulintchenko, S. Sayrac, S. Vengrova, T. Yamada-Inagawa, Random and site-specific replication termination, Methods Mol. Biol. 521 (2009) 35–53.
[6] P. Mackiewicz, A. Gierlik, M. Kowalczuk, M.R. Dudek, S. Cebra, How does replication-associated mutational pressure influence amino acid composition of proteins? Genome Res. 9 (1999) 409–416.
[7] C. Chen, C.W. Chen, Quantitative analysis of mutation and selection pressures on base composition skews in bacterial chromosomes, BMC Genomics 8 (2007) 286.
[8] N. Omont, F. Kepes, Transcription/replication collisions cause bacterial transcription units to be longer on the leading strand of replication, Bioinformatics 20 (2004) 2719–2725.
[9] G. Modiano, G. Baffistuzzi, A.G. Motulsky, Nonrandom patterns of codon usage and of nucleotide substitutions in human α-and β-globin genes: an evolutionary strategy reducing the rate of mutations with drastic effects? Proc. Natl. Acad. Sci. U. S. A. 78 (1981) 1110–1114.
[10] E.V. Barkovsky, V.V. Khrustalev, Inverse correlation between the GC content of bacterial genomes and their level of preterminal codon usage, Mol. Gen. Mikrobiol. Virusol. 24 (2009) 17–23.
[11] K.J. Fryxell, E. Zuckerkandl, Cytosine deamination plays a primary role in the evolution of mammalian isochores, Mol. Biol. Evol. 17 (2000) 1371–1383.
[12] C. Crean, Y. Uvaydov, N.E. Geacintov, V. Shafirovich, Oxidation of single-stranded oligonucleotides by carbonate radical anions: generating intrastrand cross-links between guanine and thymine bases separated by cytosines, Nucleic Acids Res. 36 (2008) 742–755.
[13] L. Gros, M.K. Saparbaev, J. Laval, Enzymology of the repair of free radicals-induced DNA damage, Oncogene 21 (2002) 8905–8925.
[14] P.J. O'Brien, T. Ellenberger, The Escherichia coli 3-methyladenine DNA glycosylase AlkA has a remarkably versatile active site, J. Biol. Chem. 279 (2004) 26876–26884.
[15] M. Kimura, The Neutral Theory of Molecular Evolution, Cambridge University Press, 1983.
[16] A.E. Darling, I. Miklós, M.A. Ragan, Dynamics of genome rearrangement in bacterial populations, PLoS Genet. 18 (2008) e1000128.
[17] S. Catakli, A. Andrieux, P. Leblond, B. Decaris, A. Dary, Spontaneous chromosome circularization and amplification of a new amplifiable unit of DNA belonging to the terminal inverted repeats in *Streptomyces ambofaciens* ATCC 23877, Arch. Microbiol. 179 (2003) 387–393.
[18] M. Imai, Y.I. Tago, M. Ihara, M. Kawata, K. Yamamoto, Role of the 5′ → 3′ exonuclease and Klenow fragment of *Escherichia coli* DNA polymerase I in base mismatch repair, MGG 278 (2007) 211–220.
[19] Y. Nakamura, T. Gojobori, T. Ikemura, Codon usage tabulated from the international DNA sequence databases: status for the year 2000, Nucleic Acids Res. 28 (2000) 292.
[20] K. Tamura, J. Dudley, M. Nei, S. Kumar, MEGA4: Molecular Evolutionary Genetics Analysis (MEGA) software version 4.0, Mol. Biol. Evol. 24 (2007) 1596–1599.
[21] V.V. Khrustalev, E.V. Barkovsky, Mutational pressure is a cause of inter- and intragenomic differences in GC-content of simplex and varicello viruses, Comput. Biol. Chem. 33 (2009) 295–302.