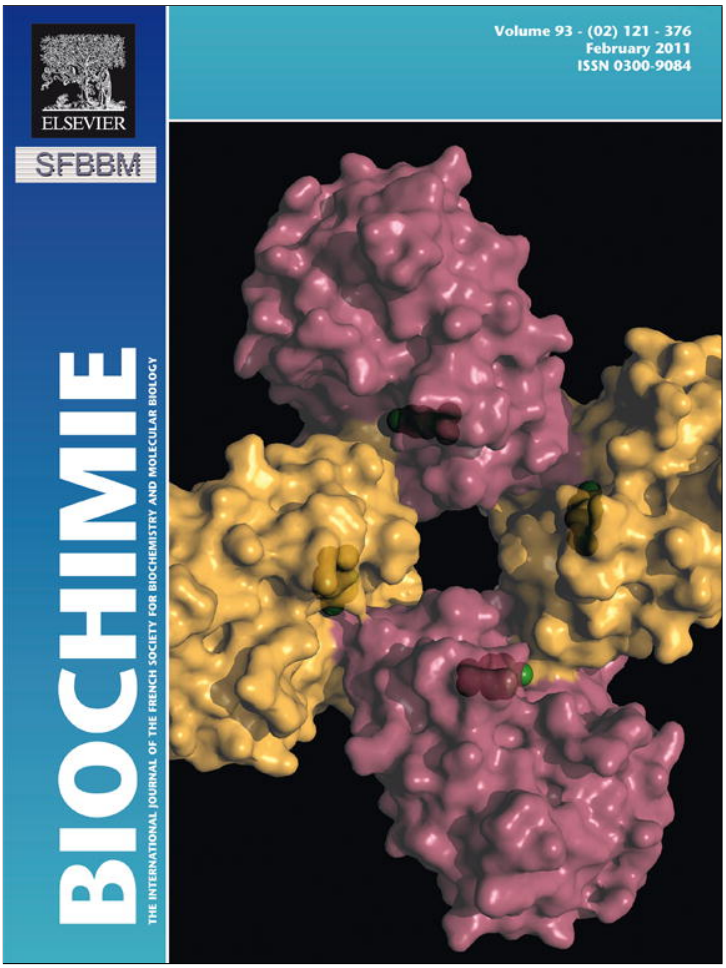

Biochimie 93 (2011) 160–167

Contents lists available at ScienceDirect

Biochimie

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

Research paper

"Protoisochores" in certain archaeal species are formed by replication-associated mutational pressure

Vladislav Victorovich Khrustalev*, Eugene Victorovich Barkovsky

Department of General Chemistry, Belarussian State Medical University, Dzerzinskogo, 83, 220116 Minsk, Belarus

ARTICLE INFO

Article history:
Received 25 May 2010
Accepted 6 September 2010
Available online 17 September 2010

Keywords:
Isochores
Replichores
Chirochores
Archaea
Mutational pressure

ABSTRACT

This report shows that isochore-like structures can be found not only in warm-blooded animals, some reptiles, fishes and yeast, but also in certain archaeal species. In perfectly shaped isochore-like structures (in "protoisochores") from *Sulfolobus acidocaldarius* and *Thermofilum pendens* genomes the difference in 3GC levels between genes from different "protoisochores" is about 30%. In these archaeal species GC-poor "protoisochores" are situated near the origin of replication, while GC-rich "protoisochores" are situated near the terminus of replication. There is a strong linear dependence between position of a gene and its 3GC level in *S. acidocaldarius* (an average difference in 3GC per 100 000 base pairs is equal to 3.6%). Detailed analyses of nucleotide usage biases in genes from leading and lagging strands led us to the suggestion that 3GC in genes situated near terminus of replication grows due to higher rates of thymine oxidation producing T to C transitions in lagging strands.

1. Introduction

Isochores are long regions of genomic DNA with relatively constant GC-content inside them. Average GC-content in different isochores is different: there are GC-rich and GC-poor isochores in human chromosomes [1]. Analogous organization of DNA has been found in chromosomes of other mammals [1], in birds [2] and even in some reptiles, fishes [3,4] and yeast [5]. It has been shown that GC-content of genes in human chromosomes correlates with time of their replication: GC-rich human isochores are usually situated near the origin of replication [6].

The most of the works on isochores are concentrated on consequences of "isochorization" [1,5,7]. For example, one of the recent studies showed that GC-rich and AT-rich chromatin domains in yeast chromosomes display distinct chromatin conformations and are marked by distinct patterns of histone modifications [5]. It has been suggested that the rate of gene duplication should be higher in "isochorized" genomes than in genomes with relatively homogenous nucleotide distribution [7]. That is why some authors suggested that isochores occurred due to positive selection [1,7]. Before the discovery of isochores in reptiles they were thought to be an adaptation for the high temperature in warm-blooded organisms of birds and mammals [1,7]. Recent studies showed that isochore-like structures are present in genomes of certain species of fish [3,4], which are cold-blooded, as well as reptiles and *Saccharomyces cerevisiae* [5].

Present work is about isochore-like structures in certain archaeal genomes (we decided to call them "protoisochores"). In our opinion, archaeal prokaryotic genomes may serve as good models for the study of GC-content heterogeneity inside a single chromosome. In model archaeal specie, *Sulfolobus acidocaldarius*, lowest levels of GC-content are characteristic for regions of DNA situated around origins of replication (OriC), while highest levels of GC-content are characteristic for regions of DNA situated strictly between two origins of replication.

Locations of OriC in chromosome of *S. acidocaldarius* have been confirmed in several experimental works [8]. So, we can state that genes situated near OriC have lower GC-content than genes situated near the region where two replication forks meet each other (Ter) in this archaeal chromosome.

An opposite situation has been described for certain bacterial species, in which GC-rich genes are usually situated near the single origin of replication and GC-poor genes are situated near the Ter region [9]. So, this article describes one more difference between bacteria and archaea. Moreover, the difference between GC-rich

Abbreviations: 3GC, usage of guanine and cytosine in third codon positions of all the 64 codons; G4f, C4f, A4f, T4f, usage of guanine, cytosine, adenine and thymine, respectively, in fourfold degenerated sites; G2f3p, C2f3p, A2f3p, T2f3p, usage of guanine, cytosine, adenine and thymine, respectively, in twofold degenerated sites from third codon positions; AA2AG, the usage of amino acid residues encoded by codons which may contain either adenine or guanine in twofold degenerated sites from their third codon positions; AA2TC, the usage of amino acid residues encoded by codons which may contain either thymine or cytosine in twofold degenerated sites from their third codon positions.

* Corresponding author. Tel.: +375 80172845957.
E-mail address: vvkhrustalev@mail.ru (V.V. Khrustalev).

0300-9084/$ – see front matter © 2010 Elsevier Masson SAS. All rights reserved.
doi:10.1016/j.biochi.2010.09.006

regions surrounding bacterial OriC and GC-poor regions around bacterial Ter [9] is much lower than the difference between GC-rich regions surrounding archaeal Ter and GC-poor regions around archaeal OriC.

Relationships between GC-content and location of a gene similar to those in chromosome of *S. acidocaldarius* have been found by us in several other completely sequenced archaeal genomes. Interestingly, in *Sulfolobus solfataricus* and *Sulfolobus tokodaii* the level of "isochorization" is much lower than in their close relative mentioned above (*S. acidocaldarius*). These facts make us suggest that replication-associated mutational pressure may form "protoisochores" in genomes of prokaryotic species only under the certain conditions.

The aim of this work was to analyze archaeal "protoisochores" and to suggest conditions under which replication-associated mutational pressure forms them.

2. Material and methods

In this work we studied 3GC levels (3GC – usage of guanine and cytosine in third codon positions of all the 64 codons; this index should not be confused with GC3 which usually refers to the GC-content in third codon positions of 56 codons (three terminal codons, single ATG codon coding for methionine and single TGG codon coding for tryptophan are excluded from GC3 calculation, as well as ATA, ATC and ATT codons coding for isoleucine)) in genes along the length of "chromosomes" from 36 completely sequenced archaeal species. Lists of codon usage for each coding region from Codon Usage Database [10] (www.kazusa.or.jp/codon) have been used as a material.

In genomes of *S. acidocaldarius*; *Thermofilum pendens*; *Methanopyrus kandleri* and *Pyrobaculum calidifontis* level of 3GC in genes shows correlation with their position in "chromosome". The main part of this article is about proper analyses of "protoisochores" and replichores found by us in the genome of *S. acidocaldarius*.

Genomes of *Aeropyrum pernix*; *Pyrobaculum arsenaticum*; *Pyrobaculum aerophilum*; *Pyrobaculum islandicum*; *Metallosphaera sedula*; *S. solfataricus* and *S. tokodaii* seem to have only "remains" of "protoisochores". These genomes have also been analyzed in this work.

In 20 archaeal genomes 3GC levels in genes are relatively equal to each other. These genomes belong to *Halobacterium sp.*; *Natronomonas pharaonis*; *Haloarcula marismortui*; *Haloquadratum walsbyi*; *Thermococcus kodakarensis*; *Methanothermobacter thermautotrophicus*; *Pyrococcus abyssi*; *Pyrococcus horikoshii*; *Pyrococcus furiosus*; *Thermoplasma volcanium*; *Thermoplasma acidophilum*; *Picrophilus torridus*; *Archaeoglobus fulgidus*; *Nanoarchaeum equitans*; *Methanococcoides burtonii*; *Methanococcus maripaludis*; *Methanococcus vannielii*; *Methanobrevibacter smithii*; *Methanococcus aeolicus*; *Methanosphaera stadtmanae*. However, some of these genomes contain short genomic islands of a different GC-content.

Although genes in genomes of *Methanocorpusculum labreanum*; *Methanospirillum hungatei*; *Methanosarcina acetivorans*; *Methanosarcina mazei* and *Methanosarcina barkeri* show wide variations in 3GC, these variations cannot be associated with replication-associated mutational pressure [11].

With the help of "Chore Viewer" algorithm, which is a modification of "Replichore Viewer" described in details in our previous work [12], we analyzed not only protoisochores but also replichores (chirochores) of archaeal genomes. This MS Excel spreadsheet available via www.barkovsky.hotmail.ru is able to build a graph showing dependence between 3GC level and position of a gene in a "chromosome" right after the pasting of all the data from "the list of codon usage for each coding region" from Codon Usage Database [10] into the special cells.

"Chore Viewer" calculates all the indexes characterizing nucleotide usage in every gene from Watson and from Crick strand separately [12]. So, one can use these indexes for comparisons between nucleotide usage in genes from leading and lagging strands.

Nucleotide usage have been calculated in fourfold degenerated sites (G4f, C4f, A4f, T4f) and in twofold degenerated sites situated in third codon positions (G2f3p, C2f3p, A2f3p, T2f3p) of every gene from the genomes studied.

Paired differences test has been applied to the data obtained. We calculated average differences between G4f and C4f, between A4f and T4f, between G2f3p and C2f3p and between A2f3p and T2f3p for genes from each strand of every replichore. Then *t*-test has been applied to each set of paired differences.

The level of nucleotide usage in twofold degenerated sites situated in third codon positions may be increased or decreased not only due to mutational pressure, but also due to the features of amino acid usage [13]. For example, in proteins of *S. acidocaldarius* the usage of amino acids encoded by codons which may contain either adenine or guanine in third codon position (AA2AG) is always higher than the usage of amino acids encoded by codons which may contain either thymine or adenine in third codon position (AA2TC). So, we calculated ratios of G2f3p and G2f3p + C2f3p in every gene and then compared them by paired differences test with ratios between AA2AG and AA2AG + AA2TC. If the ratio between G2f3p and G2f3p + C2f3p is significantly higher than the ratio between AA2AG and AA2AG + AA2TC, then the growth of G2f3p can be explained not only by the elevated usage of AA2AG, but also by asymmetric mutational pressure [13].

Differences in nucleotide usage between genes from leading and genes from lagging strands cannot be calculated directly in genomes with heterogenic GC-content distribution along the length of a chromosome. To solve this problem we compared average paired differences for genes from leading strand with average paired differences for genes from lagging strand of each replichore in *t*-test (*p*-values have been provided in text).

3. Results

3.1. Detailed study of S. acidocaldarius "protoisochores" and replichores

In Fig. 1A one can see that the usage of guanine and cytosine in third codon positions (3GC) of *S. acidocaldarius* genes strongly depends on their location. There are three origins of replication (OriC) in *S. acidocaldarius* "chromosome" [8]. Their locations have been estimated in several experimental works [8]. In Fig. 1A three OriC regions are surrounded by genes with lowest 3GC in contrast to three regions of replication termination (Ter) surrounded by genes with highest level of 3GC.

Actually, termination of replication in archaea is not site-specific, so it should take place when two replication forks meet each other [14]. In other words, replication usually stops in a region situated between two origins of replication [14].

The first peak of 3GC is situated between OriC1 and OriC2. The level of 3GC in genes near OriC1 and OriC2 is about 20%, while in genes near the first terminus of replication (Ter1) the level of 3GC is about 40%. There is a strong linear correlation between 3GC and the location of a gene between OriC1 and Ter1 (see Fig. 1A).

The slope of the linear dependence between 3GC and the location of a gene is negative between Ter1 and Oric2, although there is another peak of 3GC between Ter1 and OriC2 (see Fig. 1A).

Coefficients of correlation between 3GC and the location of a gene for region from OriC2 to Ter2 ($R = 0.55$; $p < 10^{-6}$) and for region from Ter2 to OriC3 ($R = -0.64$; $p < 10^{-6}$) are unbelievably

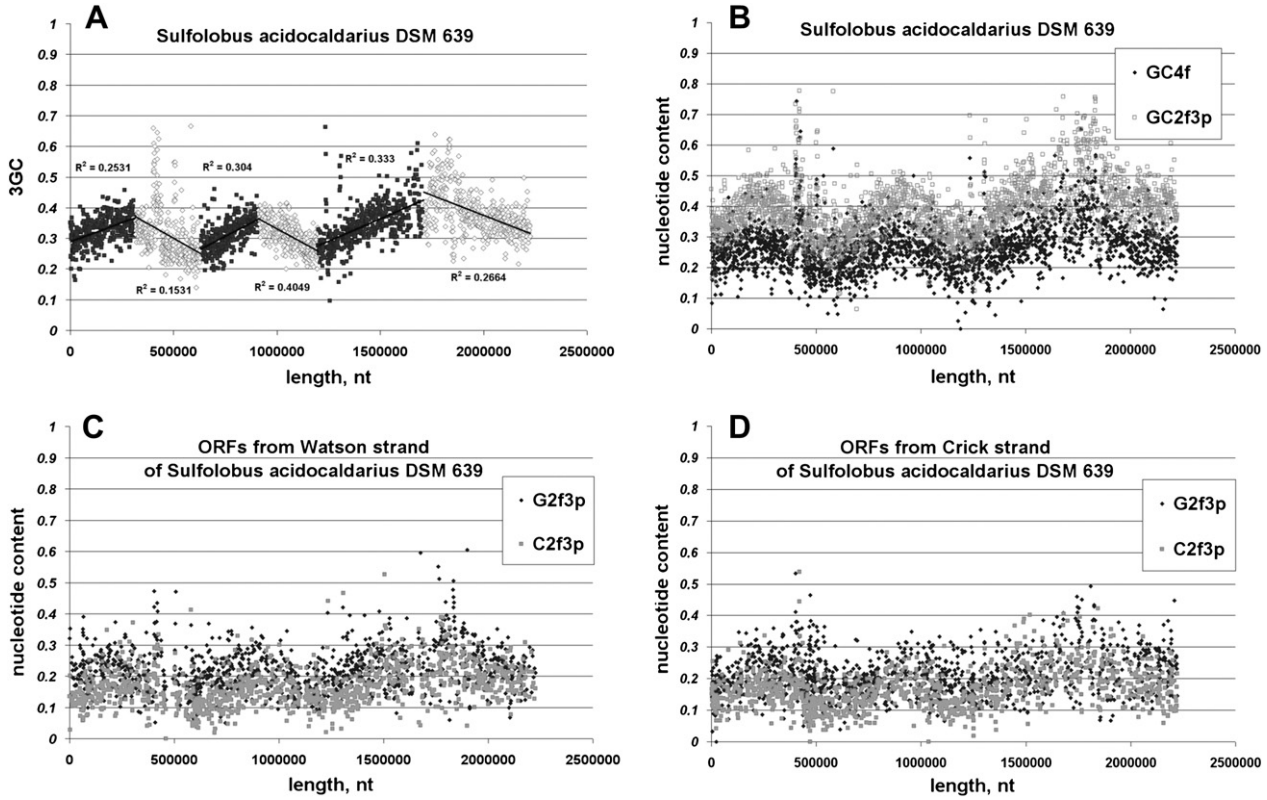

Fig. 1. Nucleotide usage in genes of *Sulfolobus acidocaldarius*. GC-content in third codon positions (3GC) and GC-content in fourfold (GC4f) and twofold degenerated sites situated in third codon positions (GC2f3p) of genes is given along the length of chromosome (A and B, respectively). Usages of guanine and cytosine in twofold degenerated sites situated in third codon positions (G2f3p and C2f3p, respectively) are given for coding regions situated on Watson strand (C) and for coding regions situated on Crick strand (D).

high and statistically significant. These dependences are not altered by genes with odd behavior. So, we can calculate slopes of these dependences. For the region between OriC2 and Ter2 the slope is the following: 3.6% of 3GC per 100 000 base pairs. For the region between Ter2 and OriC3 the slope is practically the same, but of an opposite direction: −3.6% of 3GC per 100 000 base pairs.

The distance between OriC3 and OriC1 is much longer than the distance between OriC1 and OriC2 and between OriC2 and OriC3 [8]. As one can see in Fig. 1A, the peak of 3GC in genes situated near Ter3 is higher (about 50%) than peaks of 3GC in genes situated near Ter1 and Ter2 (about 40%).

Theoretically, the level of 3GC in genes near Ter regions may be increased due to higher rates of AT to GC transitions or due to higher rates of AT to GC transversions. Fig. 1B shows that GC-content in twofold degenerated sites situated in third codon positions (GC2f3p) is always higher than GC-content in fourfold degenerated sites (GC4f) in genes from *S. acidocaldarius*.

Fig. 1C shows distribution of guanine and cytosine in twofold degenerated sites from third codon positions (G2f3p and C2f3p, respectively) in genes along the length of Watson strand from *S. acidocaldarius* "chromosome". Fig. 1D shows the distribution of the same indexes (G2f3p and C2f3p) but in genes along the length of complementary Crick strand from *S. acidocaldarius* "chromosome".

Comparison of Fig. 1C and D leads to the conclusion that genome of *S. acidocaldarius* contains six replichores (chirochores). Indeed, in region from OriC1 to Ter1 Watson strand is the leading one, while Crick strand is the lagging one. In the next region (from Ter1 to OriC2) Watson strand is the lagging one, while Crick strand is the leading one.

The difference between G2f3p and C2f3p (see Supplementary material, Tables 1 and 2) is always higher for genes from leading strands than for genes from lagging strands. This relationship between paired differences for genes from leading and lagging strands is statistically significant for four from six chirochores: 6.7% vs. 5.6% ($p = 0.228394$); 8.1% vs. 5.5% ($p = 0.009665$); 8.0% vs. 4.0% ($p = 0.000012$); 6.6% vs. 3.1% ($p = 0.000637$); 5.4% vs. 2.5% ($p = 0.002049$); 5.0% vs. 4.7% ($p = 0.732740$). The level of G2f3p is higher than C2f3p in the most of the genes from *S. acidocaldarius*, while in genes from lagging strands the difference between G2f3p and C2f3p is lower than in genes from leading strands.

The ratio between G2f3p and G2f3p + C2f3p is significantly higher ($p < 10^{-6}$) than the ratio between AA2AG and AA2AG + AA2TC for genes from leading strands. It means that the rates of A to G transitions in genes from leading strands are some higher than the rates of T to C transitions.

The ratio between G2f3p and G2f3p + C2f3p is higher than the ratio between AA2AG and AA2AG + AA2TC for genes from lagging strands only for three from six chirochores (corresponding *p*-values are: 0.001615; 0.000009 and 0.037326). In other three chirochores from *S. acidocaldarius* genome the ratio between G2f3p and G2f3p + C2f3p is equal to the ratio between AA2AG and AA2AG + AA2TC (corresponding *p*-values are: 0.069996; 0.231735 and 0.321481). It means that the rates of T to C transitions in genes from lagging strands are higher than the rates of T to C transitions in genes from leading strands.

In general, replichores of *S. acidocaldarius* behave like replichores from the most of bacterial species: the level of G is some higher in genes from leading strands than in genes from lagging strands [12,15,16].

Interestingly, there is no significant difference between the usage of guanine and the usage of cytosine in fourfold degenerated sites (G4f and C4f, respectively) of genes for four from six chirochores of

S. acidocaldarius. However, the usage of adenine and thymine in fourfold degenerated sites (A4f and T4f, respectively) of genes from leading and lagging strands of this archaeal specie differ from each other (see Fig. 2A and B).

In genes from lagging strands the level of A4f is significantly higher than the level of T4f for all the six chirochores ($p = 0.013606$; $p < 10^{-6}$; $p = 0.000002$; $p = 0.000003$; $p < 10^{-6}$; $p < 10^{-6}$). In genes from leading strands A4f is equal to T4f for four chirochores ($p = 0.384675$; $p = 0.179632$; $p = 0.15238$; $p = 0.975139$) and A4f is significantly lower than T4f for two chirochores ($p = 0.005998$; $p = 0.047152$). Analogous situation has been found for A2f3p and T2f3p distribution (Fig. 2C and D). Although A2f3p is usually higher than T2f3p (partially, because AA2AG is usually higher than AA2TC), the difference between them is higher in genes from lagging strands. For six chirochores, average difference between A2f3p and T2f3p for genes from leading strands vs. average difference for genes from lagging strands is as follows: 1.0% vs. 5.4% ($p = 0.000210$); 3.8% vs. 5.3% ($p = 0.273268$); 2.0% vs. 5.9% ($p = 0.001110$); 0.6% vs. 7.9% ($p = 0.000002$); −0.3% vs. 5.5% ($p < 10^{-6}$); −2.0% vs. 0.1% ($p = 0.027197$). Insignificant intrastrand paired differences are written in italic type.

In general, genes from lagging strands of all the six chirochores of *S. acidocaldarius* genome are relatively enriched with C2f3p, A4f and A2f3p, while genes from leading strands are relatively enriched with G2f3p, T4f and T2f3p.

3.2. Other archaeal genomes with "protoisochores" and their remains

There is a single origin of replication predicted in "chromosome" of *T. pendens* by GC-skews analysis [17]. This OriC is located in the region with relatively low GC-content. As one can see in Fig. 3A, in genes situated near the OriC of *T. pendens* "chromosome" the level of 3GC is about 60%, while in genes situated around Ter region of the same "chromosome" the level of 3GC is about 85%. The correlation between 3GC and location of a gene is strong and statistically significant ($p < 10^{-6}$) for both replichores (see Fig. 3A).

We confirmed that there are two replichores in *T. pendens* genome. There is a little excess of G4f in genes from leading strands relatively to the genes from lagging strands as well as a little excess of C4f in genes from lagging strands relatively to the genes from leading strands (see Supplementary material, Tables 3 and 4).

An average difference between AA2AG/(AA2AG + AA2TC) ratio and G2f3p/(G2f3p + C2f3p) ratio is some higher for genes from lagging strands than that for genes from leading strands (for the first chirochore $p = 0.010426$; for the second chirochore $p = 0.024174$). This fact makes us suggest that T to C transitions have higher rates of occurrence in genes from lagging strands relatively to those rates in genes from leading strands of *T. pendens* genome.

Although there is only one predicted origin of replication (OriC1) in *M. kandleri* genome (it is situated near nucleotide #1 in Fig. 3B) [18], we can suggest two other regions (see Fig. 3B) which may contain OriC (near nucleotide #400 000 and near nucleotide #800 000). In these two regions (OriC2 and OriC3) the level of 3GC reaches its lowest points.

According to our calculations, there are two well-structured replichores in *M. kandleri* genome, situated between OriC2 and Ter2 and between Ter2 and OriC3. The level of G4f is significantly higher than the level of C4f in genes from leading strands of these two replichores ($p = 0.001041$; $p < 10^{-6}$), while there is no significant difference between G4f and C4f ($p = 0.072243$; $p = 0.087010$) in genes from their lagging strands (see Supplementary material, Tables 5 and 6). The level of A4f is higher than the level of T4f in genes from leading strands of two abovementioned replichores

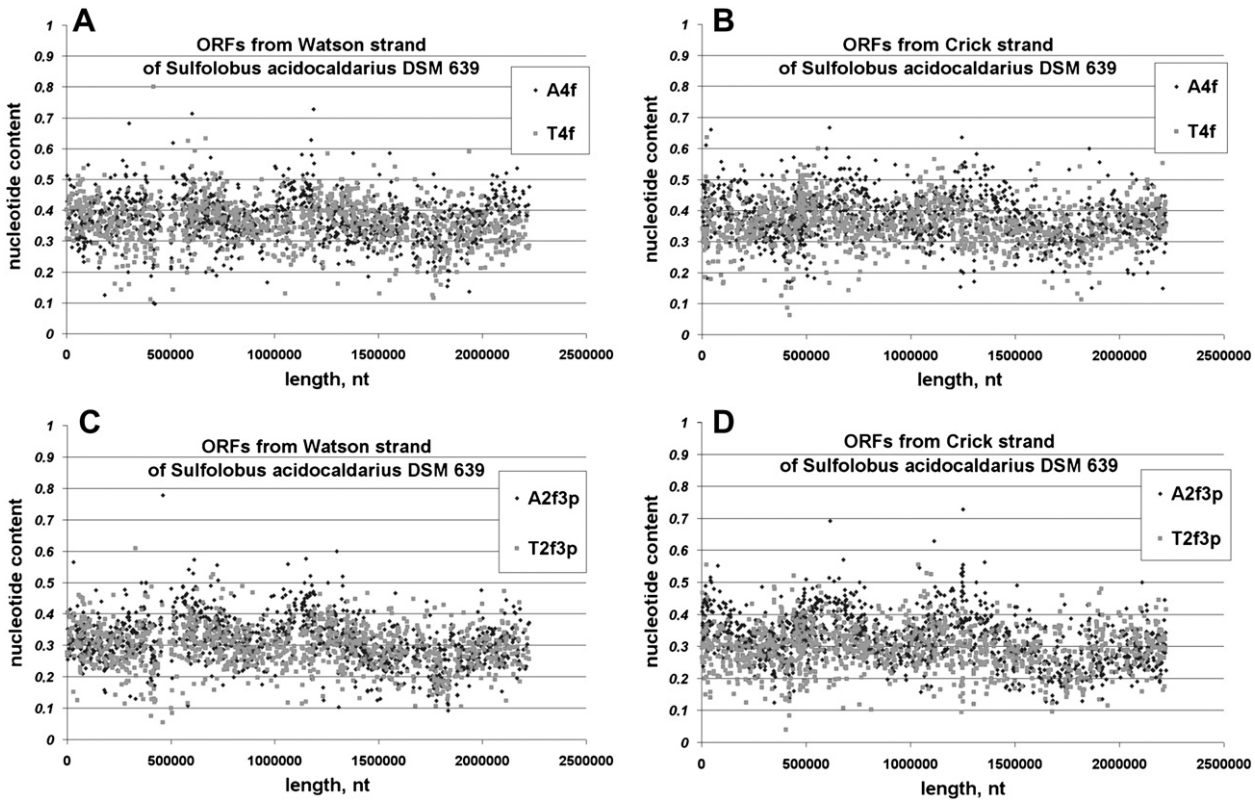

Fig. 2. Nucleotide usage in genes of *Sulfolobus acidocaldarius*. Usages of adenine and thymine in fourfold degenerated sites (A4f and T4f, respectively) are given for coding regions situated on Watson strand (A) and for coding regions situated on Crick strand (B). Usages of adenine and thymine in twofold degenerated sites situated in third codon positions (A2f3p and T2f3p, respectively) are given for coding regions situated on Watson strand (C) and for coding regions situated on Crick strand (D).

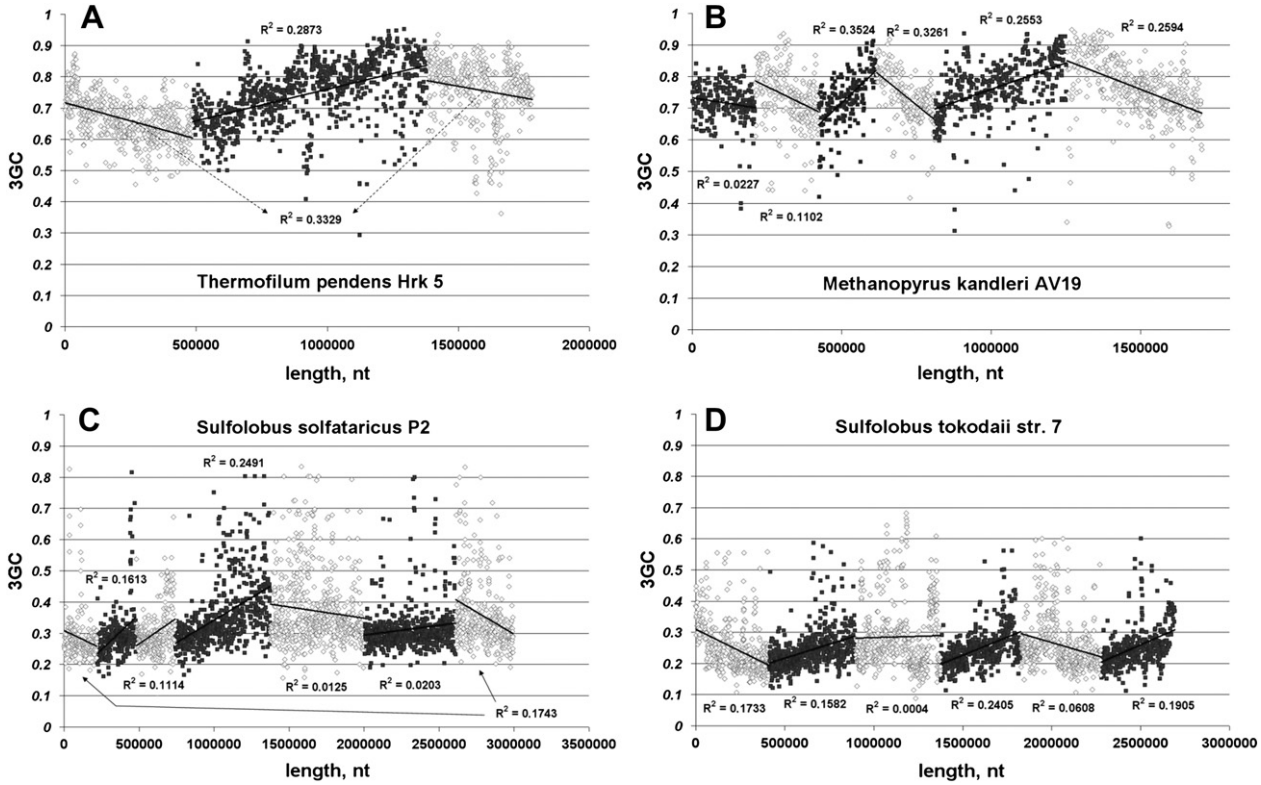

Fig. 3. GC-content in third codon positions (3GC) of genes along the length of the chromosome of *Thermofilum pendens* (A); *Methanopyrus kandleri* (B); *Sulfolobus solfataricus* (C) and *Sulfolobus tokodaii* (D).

($p = 0.012491$; $p = 0.000801$), while the level of T4f is higher than the level of A4f in genes from their lagging strands ($p < 10^{-6}$; $p = 0.015149$). There is no "chirality" in nucleotide usage biases in four other putative replichores of *M. kandleri* (see Supplementary material, Tables 5 and 6).

There are three experimentally confirmed origins of replication in *S. solfataricus* "chromosome" [8]. As one can see in Fig. 3C, the most of coefficients of correlation between position of a gene and its 3GC are low ($R < 0.5$) for *S. solfataricus*. Moreover, 3GC levels are very close to each other for the most of the genes from *S. solfataricus* genome [11]. However, it seems like several "remains" of "protoisochores" can still be seen in Fig. 3C. Namely, for genes situated in the region from OriC2 to Ter2 the coefficient of correlation between 3GC and their position is some lower than 0.5 but it is still statistically significant ($p < 10^{-6}$).

Calculations of nucleotide usage biases in *S. solfataricus* replichores showed that there is a chirality of these indexes only for replichore#1 (between OriC1 and Ter1) and replichore#2 (between Ter1 and OriC2), as well as for replichore#5 (between OriC3 and Ter3) and replichore#6 (between Ter3 and OriC1). All the biases in these replichores (see Supplementary material, Tables 7 and 8) are analogous to those described by us for replichores of *S. acidocaldarius*.

We identified three regions with minimal 3GC levels (putative regions containing OriCs) in the genome of *S. tokodaii*. As one can see in Fig. 3D, coefficients of correlation between position of a gene and its 3GC are low for *S. tokodaii*, while the "remains" of protoisochores in this genome are seen more clearly than in the genome of *S. solfataricus*. We have not found any chirality in nucleotide usage biases between genes situated on putative leading and lagging strands of *S. tokodaii* genome (see Supplementary material, Tables 9 and 10).

We also identified four regions with minimal 3GC levels in the genome of *P. calidifontis* (see Fig. 4A). Regions with highest levels of 3GC seem to be situated approximately between regions with lowest 3GC in *P. calidifontis* chromosome (see Fig. 4A). No chirality in nucleotide usage biases between genes situated on putative leading and lagging strands of *P. calidifontis* has been found by us (see Supplementary material, Tables 11 and 12).

In completely sequenced genomes of three other Pyrobaculums (*P. aerophilum*, *P. arsenaticum* and *P. islandicum*) the situation is even more complicated than in the genome of *P. calidifontis*. For example, in *P. aerophilum* genome (see Fig. 4B) one cannot find 3GC peaks, but only three regions with lower 3GC levels.

In genome of *Metallosphaera sedula* (Fig. 4C) peaks of 3GC are not situated between regions with lowest 3GC levels. In genome of *Aeropyrum pernix* one can find one region with 3GC increasing (from 50 to 85%) along the length of the "chromosome", but cannot find analogous region with 3GC decreasing along its length (see Fig. 4D).

4. Discussion

4.1. On possible molecular mechanisms of replication-associated mutational pressure in genomes of S. acidocaldarius and T. pendens

Theoretically, intragenomic differences in GC-content may occur due to incorporation of transposable elements, due to differential expression of certain genes or operons [11] as well as due to the differences in gene expression levels. In this work we showed that many intragenomic variations of GC-content in archaea occur due to replication-associated mutational pressure. In two genomes with almost perfectly shaped "protoisochores" and replichores (those of *S. acidocaldarius* and *T. pendens*) level of 3GC in genes grows linearly

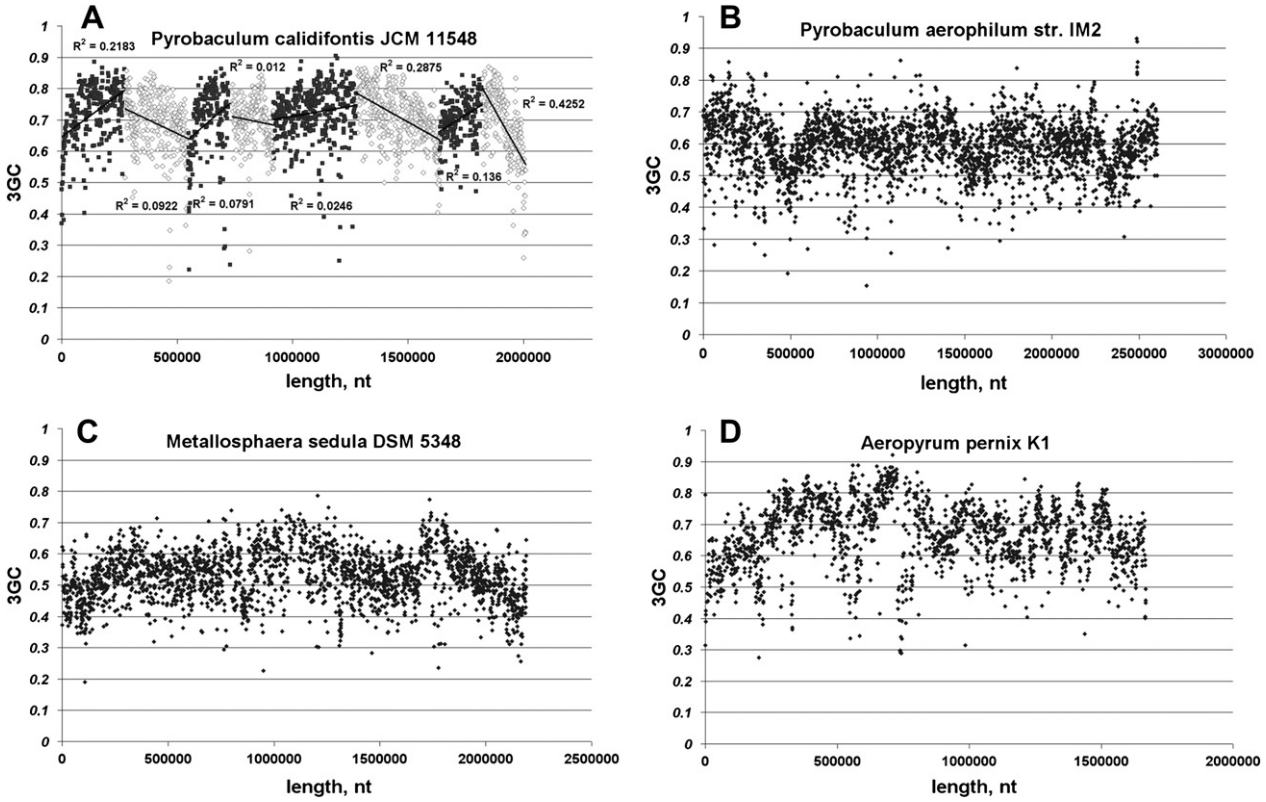

Fig. 4. GC-content in third codon positions (3GC) of genes along the length of the chromosome of *Pyrobaculum calidifontis* (A); *Pyrobaculum aerophilum* (B); *Metallosphaera sedula* (C) and *Aeropyrum pernix* (D).

with the increase of the distance from the origin of replication. Highest level of 3GC is characteristic for genes situated near the terminus of replication.

In the most of archaeal genomes with G + C lower than 60% GC-content in 2-fold degenerated sites situated in third codon positions is much higher than GC-content in 4-fold degenerated sites [11]. It makes us suggest that 3GC in genes near Ter regions of *S. acidocaldarius*, *T. pendens* and *M. kandleri* "chromosomes" is elevated due to higher rates of AT to GC transitions and not due to higher rates of AT to GC transversions.

The difference between G2f3p and C2f3p may be due to asymmetric mutational pressure or due to unequal usage of amino acids encoded by codons containing twofold degenerated sites in third positions [13]. Indeed, the usage of codons containing adenine or guanine in their twofold degenerated sites situated in third codon positions (AA2AG) is always higher than the usage of codons containing thymine or cytosine in their twofold degenerated sites situated in third codon positions (AA2TC) in *S. acidocaldarius* and *T. pendens* ORFs.

There are not only "protoisochores", but also replichores (chirochores) in genomes of *S. acidocaldarius* and *T. pendens*. This fact ensures us that origins of replication have not changed their positions in these two genomes for a long period of time [12].

However, even in chromosome of *S. acidocaldarius* a few regions with odd behavior can be found. These regions (see Fig. 1A) may be genomic islands containing integrated phage or genes translocated from regions with higher GC-content.

Replication-associated mutational pressure should be caused by higher rates of certain nucleotide mutations occurrence in single-stranded DNA than in doublestranded DNA [12]. It was shown that the rates of cytosine and adenine deamination, as well as of guanine and thymine oxidation are higher in singlestranded DNA [19]. Lagging strands exist in singlestranded form for a longer period of time than leading strands, even though the length of archaeal Okazaki's fragments is shorter than that of bacterial ones [20].

According to our calculations, the rates of T to C transitions should be higher in lagging strands of *S. acidocaldarius* genome than in leading strands. So, it is likely that the growth of 3GC in genes situated near Ter regions of this archaeal genome is also due to higher rates of T to C transitions in lagging strands.

The rates of T to C transitions in genes from *S. acidocaldarius* lagging strands should be elevated due to higher rates of thymine oxidation in singlestranded DNA. Different base-pairing features of oxidized thymine (5-formyl-uracil) have been described in different experimental works. Some researchers stated that 5-formyl-uracil in DNA template usually forms "correct" mispair with adenine, while less frequently it forms "mutagenic" mispair with guanine [21,22]. This 5-formyl-uracil:guanine mispair leads to T to C transition occurrence [21,22]. Other researchers stated that 5-formyl-uracil often forms mispairs with cytosine (leading to T to G transversions) and thymine (leading to T to A transversions), and not with guanine [23]. Moreover, experimental results showed that the probability of 5-formyl-uracil:guanine mispair formation is pH-dependent [21]. The higher is the level of pH, the higher is the probability of 5-formyl-uracil:guanine mispair formation [21].

Our results (see Figs 1 and 2) can be interpreted in the following way: 5-formyl-uracil residues occurring due to thymine oxidation in lagging strands of *S. acidocaldarius* form mispairs with both a) guanine (leading to T to C transitions) and b) with thymine (leading to T to A transitions).

5-formyl-uracil in lagging strands of *S. acidocaldarius* may also form mispairs with cytosine, but T to G transversions produced due to this process should be less frequent than G to T transversions produced by guanine oxidation [12,19,24] in lagging strands.

4.2. On the possible connection between the speed of replication fork movement during S. acidocaldarius replication and the rates of thymine oxidation in its lagging strands

The structure of *S. acidocaldarius* chirochores is practically the same near the OriC and near the Term region, although GC-content is higher near the last one. So, the excess of 3GC near the Ter should be due to the same processes of replication-associated mutational pressure, but more biased towards AT to GC transitions.

According to our hypothesis, the rates of T to C transitions in lagging strands should be higher at the end of replication than at the beginning of this process. The longer is the period of single-stranded DNA existence, the higher is the probability of thymine oxidation [19]. So, if the speed of replication fork movement is lower at the end of replication, then the rates of thymine oxidation should be higher in lagging strands near the Ter region.

According to the results of experimental work, the speed of replication fork movement is really lower at the end of *S. acidocaldarius* replication [8]. The total length of the S-phase in *S. acidocaldarius* is 96 min [8]. The distances between the origins are 630, 570, and 1020 kb. Two replication forks continue to progress for more than 40 min after the others four have terminated [8]. It means that an average speed of replication during first 56 min is equal to 11.25 kb/min (630 kb divided by 56 min). An average speed of replication during the last 40 min is equal to 9.75 kb/min (390 kb divided by 40 min). One of the possible explanations of this decline in replication speed on the way from OriC to Ter is in the decrease of the size of intracellular dNTP pool [25]. It also has been shown that increasing amount of pyrophosphate during the replication is the factor which makes the speed of replication slower (pyrophosphate is actually a product of this reaction) [26]. Special enzyme, inorganic pyrophosphatase, catalyzes hydrolysis of pyrophosphate and facilitates replication [26]. Theoretically, insufficiency of inorganic pyrophosphatase function may cause the slower rates of replication fork movement at the end of replication.

However, according to this hypothesis, not only thymine oxidation, but also cytosine and adenine deamination, as well as guanine oxidation, should increase their rates in genes situated near the Ter region. According to the data from Fig. 1, the rates of thymine oxidation are growing with the increase of singlestranded DNA existence time under the higher slope than the rates of cytosine deamination. The later effect may be due to the increase of pH at the end of DNA replication [21]. Indeed, in many eukaryotic cells (at least, in different human and Tetrahymena cells) the level of intracellular pH has been shown to grow at the end of S-phase [27,28,29]. Increase in pH level may not only modify base-pairing properties of 5-formyl-uracil inducing higher rates of 5-formyl-uracil:guanine mispair occurrence [21], but also affect the function of certain repair enzymes involved in replication-associated mutational pressure formation.

4.3. Frequent translocations of OriC regions as a factor able to cause irregularity in GC-content variations along the length of many archaeal chromosomes

In many archaeal genomes with heterogeneous GC-content distribution between genes one cannot find clearly shaped "protoisochores" and replichores. However, several "remains" of protoisochores can still be found in these genomes (see Fig. 4B, C and D). Theoretically, translocations, duplications or deletions of OriC regions should lead to this situation [12]. Translocations and inversions of relatively long regions of archaeal "chromosome" which do not contain OriC should also lead to the disturbance of the structure of both "protoisochores" and replichores [12].

However, even if the structure of "protoisochores" and replichores has been altered, mutational processes involved in formation of "protoisochores" and replichores should not be stopped. After a certain amount of generations altered archaeal chromosome should acquire "normal" structure with clearly shaped replichores and "protoisochores". According to our work, this "normal" structure of archaeal chromosome is not so common. In our opinion, translocations of OriC should be more frequent event in archaea than in bacteria, because regions of replication termination are not site-specific in archaea [14].

"Protoisochores" are present in the genome of *S. acidocaldarius*, while in genomes of two other Sulfolobuses (*S. solfataricus* and *S. tokodaii*) only "remains" of "protoisochores" can be found. Differences in 3GC between genes are much wider in *S. acidocaldarius* than in *S. solfataricus* and *S. tokodaii* [11]. It means that archaeal genomes may acquire protoisochores and may lose them in a relatively short (in evolutionary aspect) period of time. According to the selectionism-based point of view [1, 7], one should state that the structure of *S. acidocaldarius* genome is "much more progressive" than the structure of *S. solfataricus* genome. However, this seems to be an overstatement.

4.4. On several interplaying factors which may lead to the formation of "protoisochores"

There are archaeal species with relatively homogenous distribution of 3GC in their genomes (species without "protoisochores") and those with heterogeneous 3GC distribution. According to our hypothesis, several factors may lead to the formation of "protoisochores".

The decrease in speed of replication fork movement during the replication should lead to the longer period of the existence of lagging strands in form of singlestranded DNA. The size of dNTP pool in a cell before the replication may be a factor determining whether the speed of replication fork movement will be significantly lower at the end of replication [25].

Features of the singlestranded DNA reparation is also an important factor. Excision of 5-formyl-uracil from singlestranded DNA of a lagging strand during replication should prevent T to C transition occurrence [24, 30]. Probably, there is an insufficiency in this kind of repair in the most of archaeal genomes with $G + C < 60\%$. Indeed, in these genomes GC-content is growing due to AT to GC transitions, and not due to the high rates of AT to GC transversions [11].

Since base-pairing properties of 5-formyl-uracil are pH-dependent [21], the growth of pH at the end of S-phase (in case if this growth is characteristic not only for eukaryotic [26,27,28], but also for archaeal cells) may also increase the frequency of T to C transitions in genes situated near Ter regions.

In some bacterial genomes [9], as well as in eukaryotic chromosomes [6], 3GC is higher in genes situated near OriC and lower in genes situated near Ter. This pattern of 3GC structuring along "chromosome" is controversial to that found by us in certain archaeal species. Probably, 5-formyl-uracil excision from single-stranded DNA of a lagging strand is more effective in bacterial and eukaryotic species possessing isochores than in archaeal species with "protoisochores". In this case, C to U and G to 8-oxo-G mutations (frequencies of which are higher in singlestranded DNA [12,19]) should lead to the decrease of 3GC in genes situated near bacterial and eukaryotic Ter regions.

5. Conclusions

In this work we showed that unusual isochore-like GC-content heterogeneity in genomes of certain archaeal species does exist: regions of a higher GC-content are situated near the terminus of

replication, while regions of lower GC-content are situated near the origin of replication. This feature has been found only in archaeal genomes, while in bacteria [9] and in eukaryotic chromosomes [6] regions of a higher GC-content are usually situated near the origin of replication.

Our finding can be used as an additional criterion for the prediction of OriC regions location in archaeal genomes along with GC-skews method [16]. According to the present work, regions surrounded by genes with low 3GC contain origins of replication in some part of archaeal species.

In our opinion, genomes of *S. acidocaldarius* and *T. pendens* are perfect models for different kinds of *in vivo*, *in vitro* and *in silico* studies of mammalian isochores.

Appendix. Supplementary material

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.biochi.2010.09.006.

References

[1] G. Bernardi, The vertebrate genome: isochores and evolution, Mol. Biol. Evol. 10 (1993) 186–204.
[2] M. Costantini, M. Di Filippo, F. Auletta, G. Bernardi, Isochore pattern and gene distribution in the chicken genome, Gene 400 (2007) 9–15.
[3] M. Costantini, F. Auletta, G. Bernardi, Isochore patterns and gene distributions in fish genomes, Genomics 90 (2007) 364–371.
[4] C. Melodelima, C. Gautier, The GC-heterogeneity of teleost fishes, BMC Genomics 9 (2008) 632.
[5] J. Dekker, GC- and AT-rich chromatin domains differ in conformation and histone modification status and are differentially modulated by Rpd3p, Genome Biol. 8 (2007) R116.
[6] M. Costantini, G. Bernardi, Replication timing, chromosomal bands, and isochores, Proc. Natl. Acad. Sci. U.S.A. 105 (2008) 3433–3437.
[7] D.R. Forsdyke, Evolutionary Bioinformatics. Springer, New York, 2006.
[8] M. Lundgren, A. Andersson, L. Chen, P. Nilsson, R. Bernander, Three replication origins in *Sulfolobus* species: synchronous initiation of chromosome replication and asynchronous termination, Proc. Natl. Acad. Sci. U.S.A. 101 (2004) 7046–7051.
[9] V. Daubin, G. Perriere, G+C3 structuring along the genome: a common feature in prokaryotes, Mol. Biol. Evol. 20 (2003) 471–483.
[10] Y. Nakamura, et al., Codon usage tabulated from the international DNA sequence databases: status for the year 2000, Nucleic Acids Res. 28 (2000) 292.
[11] V.V. Khrustalev, E.V. Barkovsky, Study of completed archaeal genomes and proteomes: hypothesis of strong mutational AT pressure existed in their common predecessor, Genomics Proteomics Bioinformatics 8 (2010) 22–32.
[12] V.V. Khrustalev, E.V. Barkovsky, The probability of nonsense mutation caused by replication-associated mutational pressure is much higher for bacterial genes from lagging than from leading strands, Genomics 96 (2010) 173–180.
[13] V.V. Khrustalev, E.V. Barkovsky, The level of cytosine is usually much higher than the level of guanine in two-fold degenerated sites from third codon positions of genes from Simplex- and Varicelloviruses with G+C higher than 50%, J. Theor. Biol. 266 (2010) 88–98.
[14] 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.
[15] P. Mackiewicz, A. Gierlik, M. Kowalczuk, M.R. Dudek, S. Cebrat, How does replication-associated mutational pressure influence amino acid composition of proteins? Genome Res. 9 (1999) 409–416.
[16] J.R. Lobry, N. Sueoka, Asymmetric directional mutation pressures in bacteria, Genome Biol. 3 (2002) 0058.
[17] I. Anderson, J. Rodriguez, D. Susanti, I. Porat, C. Reich, L.E. Ulrich, J.G. Elkins, K. Mavromatis, A. Lykidis, E. Kim, L.S. Thompson, M. Nolan, M. Land, A. Copeland, A. Lapidus, S. Lucas, C. Detter, I.B. Zhulin, G.J. Olsen, W. Whitman, B. Mukhopadhyay, J. Bristow, N. Kyrpides, Genome sequence of *Thermofilum pendens* reveals an exceptional loss of biosynthetic pathways without genome reduction, J. Bacteriol. 190 (2008) 2957–2965.
[18] A.I. Slesarev, K.V. Mezhevaya, K.S. Makarova, N.N. Polushin, O.V. Shcherbinina, V.V. Shakhova, G.I. Belova, L. Aravind, D.A. Natale, I.B. Rogozin, R.L. Tatusov, Y.I. Wolf, K.O. Stetter, A.G. Malykh, E.V. Koonin, S.A. Kozyavkin, The complete genome of hyperthermophile*Methanopyrus kandleri AV19* and monophyly of archaeal methanogens, Proc. Natl. Acad. Sci. U.S.A. 99 (2002) 4644–4649.
[19] 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.
[20] F. Matsunaga, C. Norais, P. Forterre, H. Myllykallio, Identification of short 'eukaryotic' Okazaki fragments synthesized from a prokaryotic replication origin, EMBO Rep. 4 (2003) 154–158.
[21] A. Masaoka, H. Terato, M. Kobayashi, Y. Ohyama, H. Ide, Oxidation of thymine to 5-Formyluracil in DNA promotes misincorporation of dGMP and subsequent elongation of a mismatched primer terminus by DNA polymerase, J. Biol. Chem. 276 (2001) 16501–16510.
[22] D.E. Volk, V. Thiviyanathan, A. Somasunderam, D.G. Gorenstein, Ab initio base-pairing energies of an oxidized thymine product, 5-formyluracil, with standard DNA bases at the BSSE-free DFT and MP2 theory levels, Org. Biomol. Chem. 5 (2007) 1554–1558.
[23] H. Kamiya, N. Murata-Kamiya, N. Karino, Y. Ueno, A. Matsuda, H. Kasai, Induction of T → G and T → A transversions by 5-formyluracil in mammalian cells, Mutat. Res. 513 (2002) 213–222.
[24] L. Gros, M.K. Saparbaev, J. Laval, Enzymology of the repair of free radicals-induced DNA damage, Oncogene 21 (2002) 8905–8925.
[25] I. Odsbu, Morigen, K. Skarstad, A reduction in ribonucleotide reductase activity slows down the chromosome replication fork but does not change its localization, PLoS One 4 (2009) e7617.
[26] S.Y. Park, B. Lee, K.S. Park, Y. Chong, M.Y. Yoon, S.J. Jeon, D.E. Kim, Facilitation of polymerase chain reaction with thermostable inorganic pyrophosphatase from hyperthermophilic archaeon *Pyrococcus horikoshii*, Appl. Microbiol. Biotechnol. 85 (2010) 807–812.
[27] E. Musgrove, M. Seaman, D. Hedley, Relationship between cytoplasmic pH and proliferation during exponential growth and cellular quiescence, Exp. Cell. Res. 172 (1987) 65–75.
[28] L.K. Putney, D.L. Barber, Na-H exchange-dependent increase in intracellular pH times G2/M entry and transition, J. Biol. Chem. 278 (2003) 44645–44649.
[29] R.J. Gillies, D.W. Deamer, Intracellular pH changes during the cell cycle in Tetrahymena, J. Cell. Physiol. 100 (1979) 23–31.
[30] 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.