Available online at www.sciencedirect.com
ScienceDirect
www.sciencedirect.com/science/journal/16720229

GENOMICS PROTEOMICS & BIOINFORMATICS

Article

Study of Completed Archaeal Genomes and Proteomes: Hypothesis of Strong Mutational AT Pressure Existed in Their Common Predecessor

Vladislav V. Khrustalev* and Eugene V. Barkovsky

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

Genomics Proteomics Bioinformatics 2010 Mar; 8(1): 22-32. DOI: 10.1016/S1672-0229(10)60003-4

Abstract

The number of completely sequenced archaeal genomes has been sufficient for a large-scale bioinformatic study. We have conducted analyses for each coding region from 36 archaeal genomes using the original CGS algorithm by calculating the total GC content (G+C), GC content in first, second and third codon positions as well as in fourfold and twofold degenerated sites from third codon positions, levels of arginine codon usage (Arg2: AGA/G; Arg4: CGX), levels of amino acid usage and the entropy of amino acid content distribution. In archaeal genomes with strong GC pressure, arginine is coded preferably by GC-rich Arg4 codons, whereas in most of archaeal genomes with G+C<0.6, arginine is coded preferably by AT-rich Arg2 codons. In the genome of *Haloquadratum walsbyi*, which is closely related to GC-rich archaea, GC content has decreased mostly in third codon positions, while Arg4>>Arg2 bias still persists. Proteomes of archaeal species carry characteristic amino acid biases: levels of isoleucine and lysine are elevated, while levels of alanine, histidine, glutamine and cytosine are relatively decreased. Numerous genomic and proteomic biases observed can be explained by the hypothesis of previously existed strong mutational AT pressure in the common predecessor of all archaea.

Key words: G+C, 3GC, arginine, mutational pressure, archaea, entropy

Introduction

In this work we analyzed G+C composition of genomes and amino acid content of proteomes of all the archaeal species whose genomes have already been completely sequenced and submitted to the Codon Usage Database (www.kazusa.or.jp/codon) (*1*). We applied directed mutational pressure theory (*2, 3*) to the analysis of completely sequenced archaeal genomes.

Mutational pressure in double-stranded DNA genome is the situation when AT to GC substitution rates are not equal to GC to AT substitution rates (*2*). Most substitutions in third codon positions are synonymous, so single nucleotide mutations occurring due to mutational pressure may be fixed in these codon positions by the random genetic drift without any selective limitations (*2*).

When the level of GC content in third codon positions (3GC) for most of coding regions in the prokaryotic genome is higher than 0.5, one can suspect that this genome is under the influence of GC pressure (*4*). There should be AT pressure in the genome if 3GC level for most of its coding regions is lower than 0.5. The strongest evidence for GC pressure in ge-

*Corresponding author. E-mail: vvkhrustalev@mail.ru

nome is the situation when 3GC is higher than 1GC and 2GC for most of its genes (*4*). If 3GC level is lower than 2GC and 1GC for most of coding regions, there is AT pressure in this genome.

Mutational pressure is caused by the imbalance of mutational processes and reparation (*3, 5*). The most common and well-studied mutational processes that contribute into mutational pressure are: (i) deamination of cytosine leading to C to U transitions (*3, 5*), (ii) deamination of methylcytosine leading to 5-methyl-C to T transitions (*5, 6*), (iii) oxidation of guanine leading to G to T transversions (*3, 5*), (iv) deamination of adenine leading to A to G transitions (*5, 7*), (v) oxidation of thymine leading to T to C transitions (*5*), (vi) incorporation of 8-oxo-G into the growing DNA strand opposite adenine followed by the replacement of A with C and the excision of 8-oxo-G leading to A to C transversions (*3, 5*). Mono and polyfunctional enzymes involved in reparation of above-mentioned lesions are found in species from all three superkingdoms of life (*5-7*).

When third codon positions become saturated (due to GC pressure) or desaturated (due to AT pressure) with G and C, the probability of substitutions occurring in first and second codon positions increases (*4*). This situation has been called a strong mutational pressure (*8*) that leads to the simplification of the amino acid content of proteome (*9*).

In the work on amino acid biases of halobacterial proteomes, an elevated level of aspartic acid (which is higher than that predicted according to dinucleotide composition of genomic DNA) has been found (*10*). Authors speculated that increased level of Ala and decreased level of Lys in halobacterial proteomes are “dragging effects” caused by the compositional shift of halobacterial DNA, which would have changed to increase principally the fraction of aspartic acid alone (*10*). Another study on the haloarchaeal secretomes showed that frequencies of Lys, Ile and Leu are much lower, but the frequencies of Arg, Thr, Val and Gly are higher than those in bacterial signal peptides (*11*).

In our work we found out that amino acid biases in archaeal proteomes, as well as genomic biases characteristic to most of archaeal species, can be simply explained by a single hypothesis of strong mutational AT pressure existed in the common predecessor of all archaea.

Results and Discussion

Estimation of the strength of mutational pressure in archaeal genomes

The estimation of the strength of mutational pressure in archaeal genomes is shown in **Figure 1**. Mutational pressure is strong if the module of the coefficient of correlation (R) between GC content in third codon positions (3GC) and total GC content (G+C) for all coding regions in the genome (excluding those from genomic islands) is lower than 0.5 (*8*). Strong mutational GC pressure leads to the almost equivalently high saturation of third codon positions with G and C in all coding regions (Figure 1B) (*4, 8*). Strong mutational AT pressure leads to the almost equivalent desaturation of third codon positions with G and C through the coding genome (Figure 1D) (*8*).

Usually the picture of GC content distribution in all genes from archaeal genome can be described as “head and tail” (Figure 1A, B and D), where a head is formed by most of the coding regions with close GC content and a tail is formed by coding regions from genomic islands. Genome of *Methanosarcina acetivorans* (Figure 1C) is an exceptional one: you cannot find any “head”, but just a long “tail” that lasts from G+C = 0.24 to G+C = 0.60.

Even in the genome of *Halobacterium sp.* NRC-1 with total GC content about 0.68, there is a strong correlation between 3GC and G+C (R=0.864). As one can see in Figure 1B, this correlation is due to the small part of coding regions with lower GC content, while most of the genes (the “head” forming a kind of “roof”) have 3GC levels about 0.87. However, for coding regions with G+C>0.62 the correlation is already low (R=0.489). It means that there is strong mutational GC pressure in the main part of *Halobacterium sp.* genome (which includes 77.7% of coding regions), but there is no strong mutational GC pressure in the *Halobacterium sp.* genomic islands. Main parts of *Natronomonas pharaonis* and *Haloarcula marismortui* genomes (genes with G+C>0.60) are also under the influence of strong mutational GC pressure.

The coefficient of correlation between 3GC and G+C for *Sulfolobus tokodai* is 0.716. If we calculate this coefficient only for coding regions with G+C<0.4

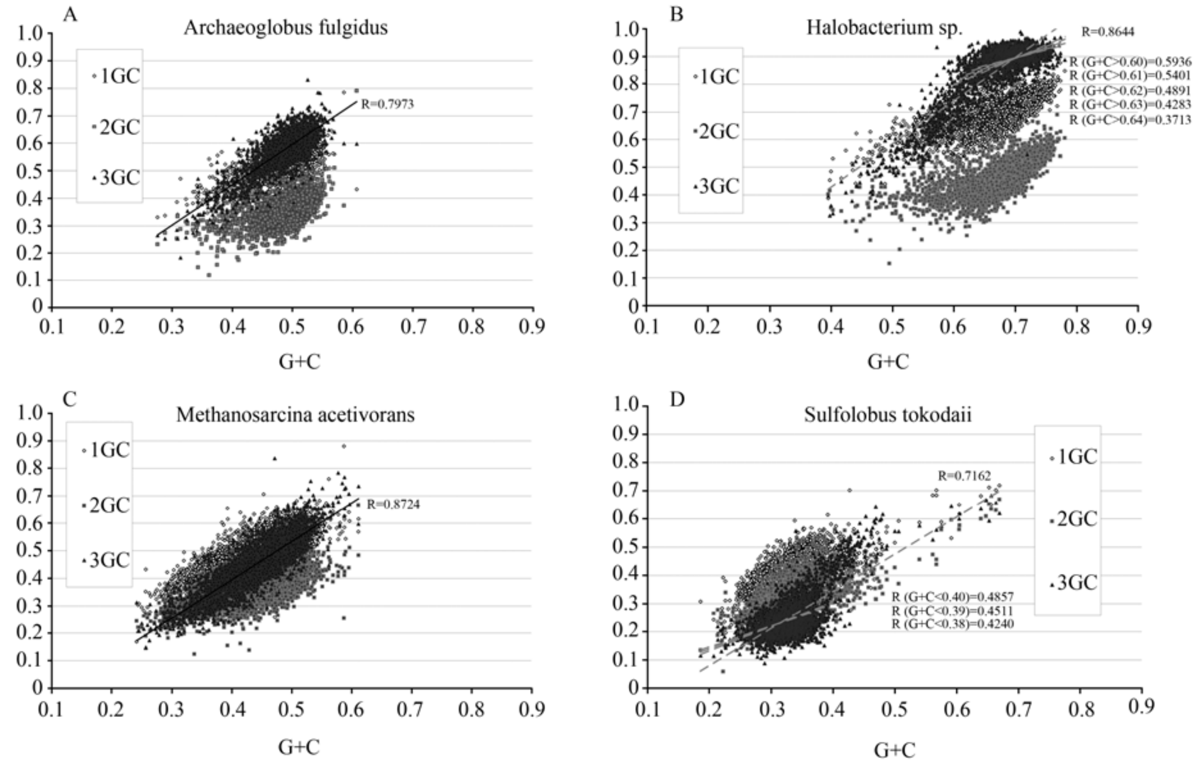

Figure 1 Intragenomic dependences between total GC content (G+C) and GC content in three codon positions reflecting the strength and direction of mutational pressure in certain archaeal coding genomes.

(Figure 1D), the situation characteristic for the strong AT pressure (low correlation between 3GC and G+C) will be found (R=0.486). There are about 5.9% of *S. tokodai* coding regions with G+C more than 0.4. Strong mutational AT pressure has also been found in the main part of *Sulfolobus solfataricus* coding genome (in genes with G+C<0.4) and in whole genomes of *Nanoarchaeum equitans*, *Methanobrevibacter smithii*, *Methanococcus aeolicus* and *Methanosphaera stadtmanae*.

Bias in Arg4 and Arg2 usage correlates with bias in GC4f and GC2f3p: genomic indicators of previously existed strong AT pressure

For the levels of arginine codon usage, it is expectable that GC-rich Arg4 codons (CGX) are used with a great preference in GC-rich genomes (**Figure 2**). In genomes with strong mutational AT pressure, there is also an expectable bias in Arg4/Arg2 codon usage: GC-poor Arg2 codons (AGA/G) are used with the great preference (Figure 2).

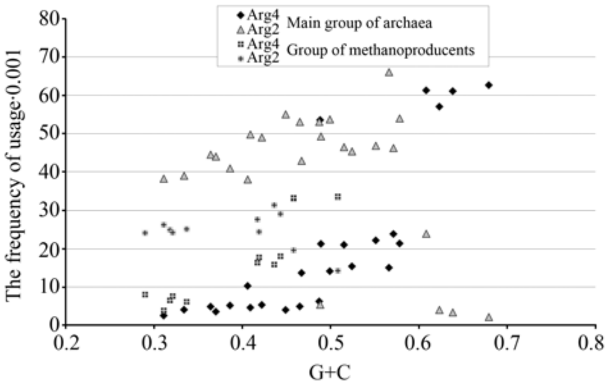

Figure 2 Dependences between average genomic GC content (G+C) and levels of arginine codon usage (Arg2 and Arg4) in archaeal genomes.

The strange is the fact that there is a great bias in Arg2 and Arg4 codon usage in the absence of strong mutational pressure. In most of archaeal genomes with average GC content, Arg2 codons are used with the great preference against Arg4 codons (Figure 2). The level of Arg4 usage is increasing with the growth of GC content but even in *Thermofilum pendens* (G+C=0.58) it is more than twofold lower than the level of Arg2 usage. Only in the genome of *Methanopyrus kandleri* with G+C=0.61 the bias became controversial: Arg4>Arg2. However, the level of *M. kandleri* Arg2 usage is still high in relation to the archaeal genomes with strong GC pressure. High level of AT-rich Arg2 codons and low level of GC-rich Arg4 codons in archaea with average GC content should be the "trace" of previously existed strong AT pressure in their common predecessor.

As one can see in Figure 2, the bias in Arg4 and Arg2 codon usage still exists in methane-producing archaea with G+C<0.46, but the level of Arg2 is some lower, while the level of Arg4 is some higher in them in comparison with genomes from the main archaeal group. It is important to highlight that two methane-producing archaea (*M. kandleri* and *Methanothermobacter thermautotrophicus*) are classified into the main group: their biases in arginine codon usage and amino acid content are different from biases found in their relatives.

In Figure 2 one can see outlying points with inversed Arg4/Arg2 bias that belong to *Haloquadratum walsbyi*. As shown in **Table 1**, *H. walsbyi* has an elevated level of 1GC relative to the species with the same total GC content (around 0.5). On the other hand, the level of 3GC in *H. walsbyi* is too low for "normal" archaea with average G+C.

H. walsbyi is phylogenetically closely related to GC-rich *Halobacterium sp.*, *H. marismortui* and *N. pharaonis* (*12-14*), genomes of which are under the influence of strong mutational GC pressure. However, we can describe mutational pressure in *H. walsbyi* genome as the weak AT pressure. We came to the conclusion that the direction of mutational pressure in the genome of *H. walsbyi* has changed not very long time ago. There was a strong GC pressure in the

Table 1 GC content and levels of arginine codon usage for genomes from the main group of archaea

Species	G+C	1GC	2GC	3GC	GC4f	GC2f3p	Arg4 (×0.001)	Arg2 (×0.001)
Halobacterium sp. NRC-1	0.679	0.698	0.466	0.872	*0.882*	*0.844*	62.68	2.13
Natronomonas pharaonis	0.638	0.694	0.444	0.775	*0.791*	*0.742*	61.07	3.33
Haloarcula marismortui	0.623	0.673	0.440	0.757	*0.745*	*0.720*	57.01	4.02
Haloquadratum walsbyi	**0.488**	**0.613**	**0.430**	***0.421***	**0.404**	**0.412**	**53.54**	**5.39**
Methanopyrus kandleri	0.608	0.656	0.420	0.749	0.734	0.761	61.28	23.92
Thermofilum pendens	0.578	0.596	0.412	0.724	0.678	0.803	21.39	53.94
Pyrobaculum calidifontis	0.571	0.594	0.415	0.705	0.703	0.738	23.90	46.21
Aeropyrum pernix	0.566	0.592	0.437	0.669	0.616	0.778	15.08	66.04
Pyrobaculum arsenaticum	0.551	0.578	0.408	0.668	0.667	0.711	22.18	46.78
Thermococcus kodakarensis	0.524	0.554	0.363	0.653	0.615	0.716	15.43	45.34
Pyrobaculum aerophilum	0.515	0.550	0.403	0.592	*0.623*	*0.586*	21.06	46.43
Methanothermobacter thermautotrophicus	0.499	0.546	0.391	0.559	0.509	0.629	14.13	53.70
Pyrobaculum islandicum	0.489	0.552	0.397	0.520	0.519	0.529	21.28	49.22
Archaeoglobus fulgidus	0.487	0.525	0.360	0.577	0.512	0.652	6.25	53.03
Thermoplasma acidophilum	0.467	0.490	0.370	0.541	0.524	0.582	13.67	42.83
Metallosphaera sedula	0.465	0.500	0.366	0.528	0.462	0.600	4.92	53.03
Pyrococcus abyssi	0.449	0.500	0.346	0.501	0.417	0.597	3.91	55.00
Pyrococcus horikoshii	0.422	0.473	0.365	0.430	0.367	0.486	5.33	48.96
Pyrococcus furiosus	0.409	0.490	0.342	0.394	0.321	0.455	4.60	49.73
Thermoplasma volcanium	0.406	0.462	0.348	0.406	0.358	0.456	10.30	37.96
Picrophilus torridus	0.368	0.415	0.325	0.365	0.313	0.418	5.18	40.86
Sulfolobus acidocaldarius	0.370	0.443	0.333	0.335	0.265	0.387	3.53	43.95
Sulfolobus solfataricus	0.364	0.429	0.330	0.334	0.270	0.376	4.86	44.47
Sulfolobus tokodaii	0.334	0.422	0.324	0.256	0.200	0.282	4.04	38.93
Nanoarchaeum equitans	0.311	0.396	0.284	0.252	0.237	0.261	2.54	38.20

genome of *H. walsbyi* predecessor, the level of 1GC was high and the bias between Arg4 and Arg2 (Arg4>>Arg2) was great.

The product of gene *MutY* is DNA glycosylase that excises A opposite G, C and 8-oxo-G (*3, 5*). This enzyme is involved in A to C and T to G transversion mechanism. Probably, some amino acid substitutions had altered the function of *H. walsbyi* MutY protein or decreased its specificity for 8-oxo-G residues and the change of mutational pressure direction has occurred.

GC4f is the GC content in fourfold degenerated sites where all nucleotide substitutions are synonymous. GC2f3p is the GC content in twofold degenerated sites from third codon positions where only transitions are synonymous. As shown in Table 1, in all archaeal species, except those that are under strong mutational GC pressure and *Pyrobaculum aerophilum*, GC4f is higher than GC2f3p. In most of them the difference between GC4f and GC2f3p is great enough to state that the rates of GC to AT transversions in archaeal genomes are higher than the rates of AT to GC transversions, but the rates of GC to AT transitions are lower than the rates of AT to GC transitions. The level of Arg4 cannot be elevated in most of archaeal species because the rates of AT to GC transversions are much less frequent than the rates of GC to AT transversions.

In *P. aerophilum* GC4f (0.623) is a little higher than GC2f3p (0.586), but the bias in arginine codons is the same as that for the most part of archaea. This may be explained by the suggestion that the bias in arginine codon usage is a much more “retrospective” indicator of previously existed mutational pressure than the bias in GC4f and GC2f3p.

In **Table 2** one can see that the bias in GC4f and GC2f3p is lower for Methanosarcina species than for the main group of archaea. In genomes of *Methanocorpusculum labreanum* and *Methanospirillum hungatei* the level of GC4f is much higher than the level of GC2f3p. Increased rates of AT to GC transversions in these genomes have already resulted in the change of arginine codon usage bias (Arg4>Arg2).

One can speculate that the bias in arginine codon usage is due to the increased (or decreased) number of tRNA clones recognizing Arg4 (or Arg2) codons. However, this alternative hypothesis is not working well: different biases in arginine codon usage are observed in species with the same number of tRNA copies; close biases are found in species with different numbers of tRNA copies (**Table 3**).

Table 2 GC content and levels of arginine codon usage for genomes from the group of methane-producing archaea

Species	G+C	1GC	2GC	3GC	GC4f	GC2f3p	Arg4 (×0.001)	Arg2 (×0.001)
Methanocorpusculum labreanum	0.508	0.558	0.396	0.571	*0.600*	*0.480*	33.53	14.24
Methanospirillum hungatei	0.458	0.539	0.388	0.447	*0.461*	*0.390*	33.18	19.60
Methanosarcina acetivorans	0.443	0.508	0.366	0.454	*0.449*	*0.437*	18.04	29.09
Methanosarcina mazei	0.436	0.514	0.365	0.428	0.413	0.427	15.86	31.43
Methanococcoides burtonii	0.419	0.506	0.356	0.394	0.341	0.408	17.73	24.44
Methanosarcina barkeri	0.417	0.501	0.361	0.391	0.368	0.389	16.39	27.70
Methanococcus maripaludis	0.337	0.443	0.317	0.252	0.206	0.255	6.14	25.16
Methanococcus vannielii	0.321	0.427	0.311	0.224	0.195	0.223	7.59	24.24
Methanobrevibacter smithii	0.318	0.432	0.317	0.206	0.150	0.210	6.48	24.95
Methanococcus aeolicus	0.311	0.407	0.306	0.222	*0.231*	*0.202*	3.86	26.28
Methanosphaera stadtmanae	0.290	0.423	0.312	0.134	0.074	0.140	8.00	24.13

Table 3 Number of tRNA clones recognizing arginine codons and levels of their usage for certain archaeal genomes

Species	Number of tRNA clones recognizing the following codons coding for arginine						Arg2 (×0.001)	Arg4 (×0.001)
	Arg2		Arg4					
	AGA	AGG	CGA	CGU	CGG	CGC		
Haloarcula marismortui	1	1	1	–	1	1	4.02	57.01
Haloquadratum walsbyi	1	1	1	–	–	1	5.39	53.54
Archaeoglobus fulgidus	1	1	1	–	1	1	53.03	6.25
Pyrococcus abyssi	1	1	1	–	1	1	55.00	3.91
Sulfolobus tokodaii	1	1	1	–	1	1	38.93	4.04

In addition, one should remember that the level of Arg4 codon usage makes a great contribution to the total level of CpG dinucleotide usage (*15*).

Mosaic structure of *M. acetivorans* genome

Figure 3 shows the distribution of 3GC in coding regions of two genomes. Several genomic islands with higher 3GC are clearly seen in the genome of *H. walsbyi* (Figure 3A), but 3GC levels of "normal" coding regions do not vary widely (from 0.33 to 0.43, approximately). Figure 3B shows that the genome of *M. acetivorans* consists of numerous short genomic islands significantly different in their GC content. In some of those "microislands", mutational pressure should have AT to GC direction (3GC higher than 50%), in others the direction of mutational pressure should be different (AT pressure). Wide variation in 3GC levels (from 0.2 to 0.8) in *M. acetivorans* coding regions is close to that in eukaryotic chromosomes (*6*).

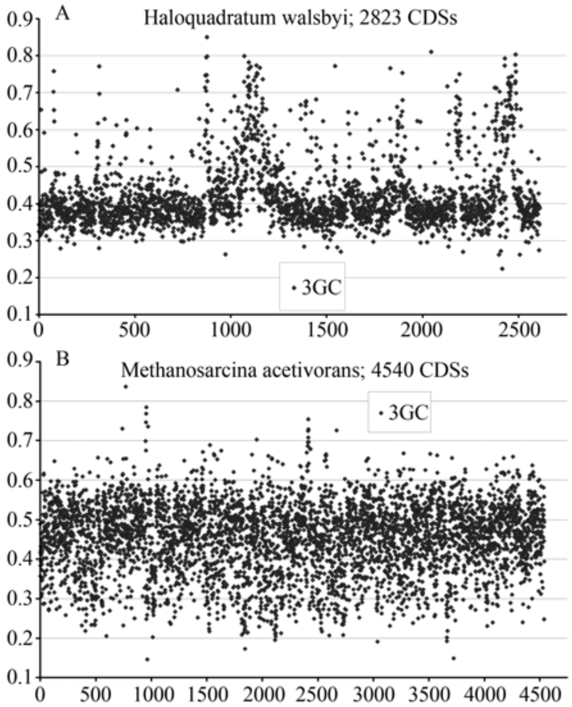

Figure 3 Distribution of GC content in third codon positions (3GC) of coding regions along the length of *Haloquadratum walsbyi* (A) and *Methanosarcina acetivorans* (B) genomes.

Methanosarcina species are known to have different stages in their life cycles (*16*). *M. acetivorans* can live as separate cells, as a cell lining and as a multicellular conglomerate with differentiated cells (*17*). In different cell types different genes are expressed. If mutator-gene is expressed only in a given stage of a life cycle, it will cause nucleotide substitutions mostly in genes that are also expressed in this differentiated cell. If numerous genes are not translated in the differentiated cell, mutator-gene will rarely cause nucleotide substitutions in them.

Coding regions might "jump" from GC-rich "microislands" to GC-poor ones and *vise versa* in *M. acetivorans*. This process should result in the absence of the great bias in arginine codon usage as well as in the growth of entropy of amino acid content distribution.

Analogous mosaic genome structure is also a characteristic of *Methanosarcina mazei* and *M. hungatei* genomes. In *Methanococcoides burtonii* and *Methanosarcina barkeri* genomes, variations in 3GC are not so wide, but they are still wider than in genomes from the main group of archaea. In *Methanococcus maripaludis*, *Methanococcus vannielii*, *Methanobrevibacter smithii*, *Methanococcus aeolicus* and *Methanosphaera stadtmanae*, the distribution of 3GC is identical to that in the main group of archaea; there are even no genomic islands in them, but the bias in arginine codon usage and specific amino acid content features are similar to those of *M. acetivorans* (Figure 2 and Table 2). The GC content of *M. hungatei* and *M. labreanum* genomes is higher than that of *M. acetivorans*; the size of their genomic islands is larger than in *M. acetivorans*, while the variations in 3GC are not so wide.

Probably, common predecessor of the group of methane-producing archaea existed for a long time with the state of genome organization similar to that of *M. acetivorans*, but then one group of its offspring drifted to AT pressure and another group drifted to GC pressure, loosing their mosaic G+C structure that remains in *M. acetivorans*.

Entropy of amino acid content distribution in archaeal genomes

We calculated entropy of amino acid content distribution (according to Claude Shannon's information theory) in all proteins coded by genomes of the main group of archaea and in all proteins from the group of methane-producing archaeal species. Entropy (the quantity of information) is the measure of uncertainty and diversity of any biological system (*9*). The lower the level of entropy, the higher the level of amino acid content uniformity.

In general, **Figure 4** shows that both GC and AT pressure lead to the decrease in entropy of amino acid content distribution. GC pressure causes increase in levels of four amino acids coded by GC-rich codons (GARP) and decrease in levels of six amino acids coded by AT-rich codons (FYMINK). On the contrary, AT pressure causes increase in levels of FYMINK and decrease in GARP usage.

The entropy of amino acid content distribution is significantly higher in proteins from the group of methane-producing archaeal species (Figure 4B) coded by genes with G+C from 0.3 to 0.6 than in proteins from the main group of archaea coded by genes with the same GC content (Figure 4A).

The highest entropy of the main group of archaea is in proteins coded by genes with G+C from 0.4 to 0.5 (Figure 4A), while the highest entropy of the group of methane-producing archaeal species is in proteins coded by genes with G+C from 0.5 to 0.6 (Figure 4B). Figure 4C shows the cause of the slow (yet statistically significant) decrease in entropy under the influence of GC pressure in proteins of the main group of archaea: the total level of 10 a.a. usage (the usage of 10 amino acids coded by codons with average GC content) increases with the growth of GC content up to the point of G+C = 0.7 and begins to decrease only in proteins coded by genes with G+C higher than 0.7. In comparison, Figure 4D shows that in proteins from the group of methane-producing archaeal species, the level of 10 a.a. usage begins to decrease in proteins coded by genes with G+C higher than 0.5.

Entropy falls more steeply in GC-poor genes (G+C<0.3) from the group of methanoproducents (Figure 4B) than in the main group of archaea (Figure 4A). This fact is surely caused by the absence of statistically significant difference between levels of 10 a.a. in proteins coded by genes with G+C from 0.2 to 0.3 and genes with G+C from 0.3 to 0.4 in the main group of archaea.

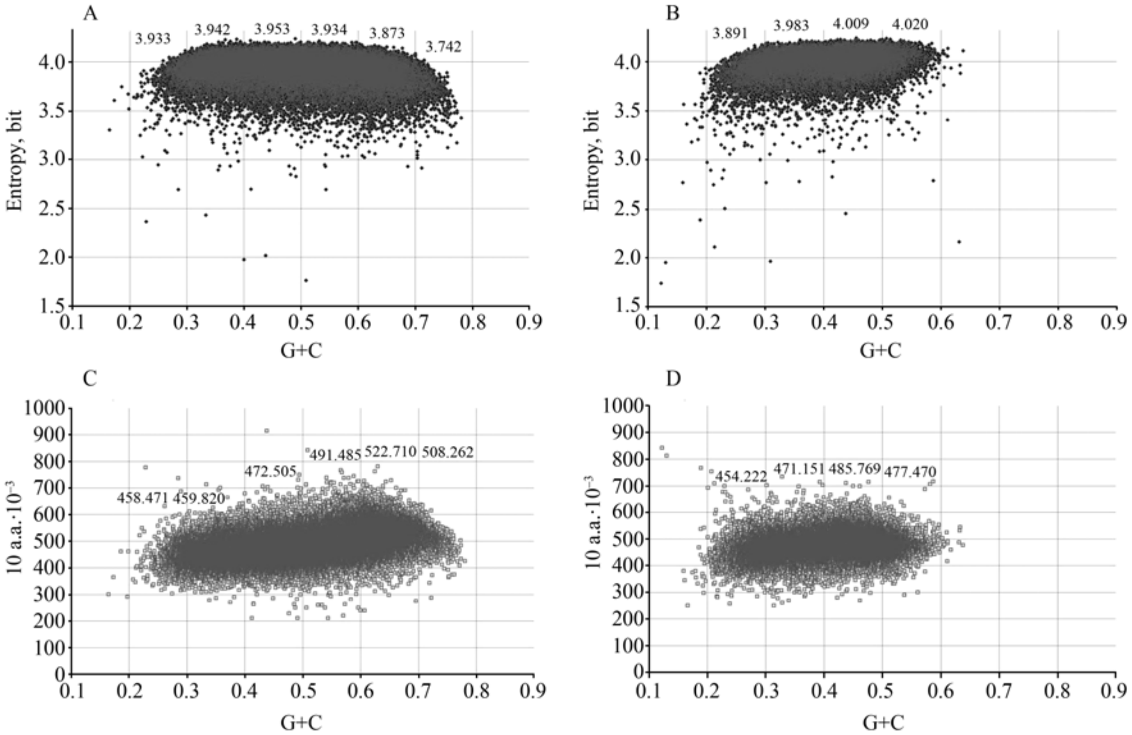

Figure 4 Dependences on GC content of the entropy of amino acid usage distribution (A, B) and the level of 10 a.a. usage (C, D) in proteins from the main group of archaea (A, C) and in proteins from the group of methane-producing archaea (B, D). 10 a.a. is the total level of usage for ten amino acids coded by codons average in GC content.

Amino acid usage in two groups of archaeal species: proteomic indicators of previously existed strong AT pressure

The levels of amino acid usage in proteins from the main group of archaea is shown in **Figure 5**. Levels of valine, threonine, aspartic acid, histidine and glutamine (Figure 5C) increase with the growth of GC content just like levels of GARP amino acids do (Figure 5A). Only the level of glutamine begins to decrease in proteins coded by G+C >0.7, when levels of other four amino acids keep growing. To find out the cause of this unexpected growth, we analyzed behavior of its possible sources (FYMINK amino acids).

One can see in Figure 5B that levels of isoleucine and lysine are much higher than the level of any other amino acid from FYMINK group in proteins coded by genes with G+C<0.6 (*18*). Great shift in lysine usage can be seen between proteins coded by genes with G+C from 0.5 to 0.6 and proteins coded by genes with G+C from 0.6 to 0.7. The greatest shift in aspartic acid usage (Figure 5C) has also taken place between proteins coded by genes with G+C from 0.5 to 0.6 and proteins coded by genes with G+C from 0.6 to 0.7. We supposed that the level of aspartic acid grew in proteins coded by genes with G+C from 0.6 to 0.7 due to the decrease in lysine (*18*). Indeed, aspartic acid is coded by GAT and GAC codons, while lysine is coded by AAA and AAG. The easiest pathway of Lys to Asp substitution is two-step nucleotide mutation AAA to GAC. This kind of two-step mutation should be frequent when mutational GC pressure is caused by both AT to GC transitions and AT to GC transversions. Increase in glutamine and histidine levels can also be due to the decrease in lysine level under the influence of GC pressure. Decrease in isoleucine should give source to the increase in valine and in threonine under the influence of GC pressure.

The level of alanine is growing more steeply than levels of other three amino acids from the GARP group under the influence of strong GC pressure (Figure 5A). Amino acid substitutions leading to alanine appearance should be more neutral than amino acids leading to glycine, arginine or proline appearance (*9*). We hypothesize that isoleucine and lysine levels are growing so steeply under the influence of

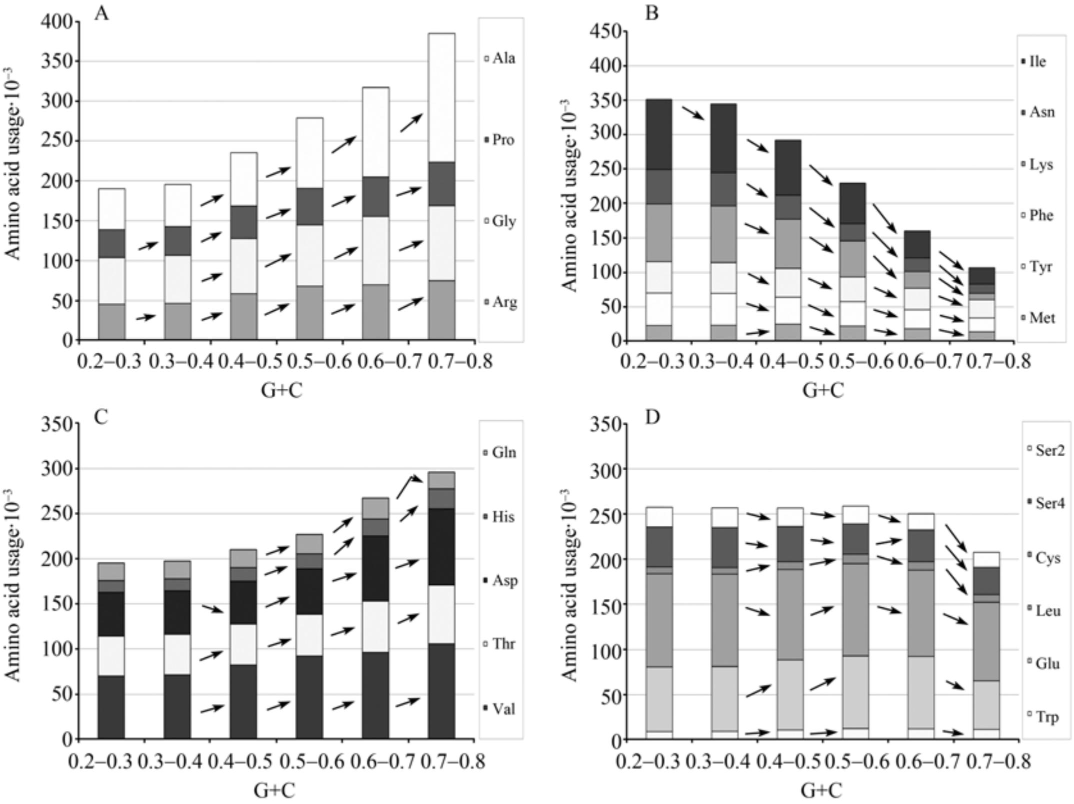

Figure 5 Levels of amino acid usage in proteins from the main group of archaea.

AT pressure because of the same circumstances: substitutions leading to isoleucine and lysine appearance should be more neutral than substitutions leading to phenylalanine, tyrosine, methionine or asparagine appearance.

As shown in Figure 5, there are only three amino acids (Ile, Pro and Arg) whose levels are significantly different between groups of proteins coded by genes with G+C from 0.2 to 0.3 and proteins coded by genes with G+C from 0.3 to 0.4. In contrast, in proteins from the group of methane-producing archaea, levels of GARP amino acids decrease, levels of FYMINK amino acids (except methionine) increase and levels of eight out of ten amino acids coded by codons with average GC content do decrease under the influence of strong AT pressure (**Figure 6**).

The absence of difference between amino acid content of archaeal proteins coded by genes with G+C from 0.2 to 0.3 and proteins coded by genes with G+C from 0.3 to 0.4 is the strongest proteomic evidence of our hypothesis. This situation could be possible only if archaeal species from the main group "came back" from the strong AT pressure. The higher rates of AT to GC transitions than in their common predecessor have not led to the significant rearrangements in amino acid content, just like it happened with *H. walsbyi* that "came back" from GC pressure.

Relatively elevated levels of histidine, cysteine and glutamine are found in proteomes that belong to archaeal species from the group of methanoproducents. This feature should be the cause of higher levels of entropy in their proteins. The level of lysine is much lower than that in the main group of archaea for proteins coded by genes with G+C from 0.4 to 0.5. This may be caused by higher rates of AT to GC transversions in genomes of mentioned methanoproducents: level of lysine has already declined, giving substrate for the growth of histidine and glutamine, and the level of aspartic acid is not growing under the influence of GC pressure in proteins coded by genes with G+C from 0.5 to 0.6. The level of isoleucine has not decreased in proteins coded by genes with G+C from 0.4 to 0.5 just like the level of lysine has done, and so valine and threonine levels keep on growing in proteins coded by genes with G+C from 0.5 to 0.6.

Mosaic genome structure should be the cause of amino acid content diversification in the common predecessor of arachaea from the group of methanoproducents. However, relatively elevated levels of isoleucine and lysine persist in this group of archaea, providing the evidence of previously existed strong AT pressure in their common predecessor.

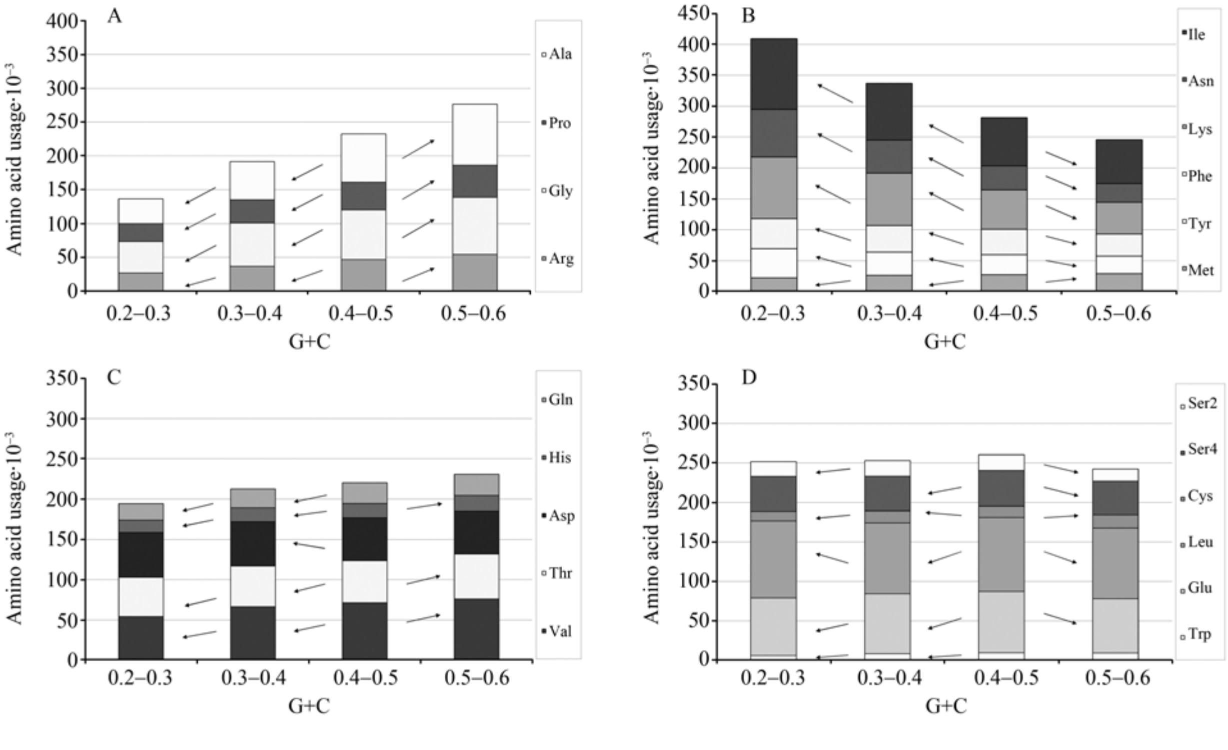

Figure 6 Levels of amino acid usage in proteins from the group of methane-producing archaea.

Conclusion

All the genomic and proteomic data obtained in our research can be explained by the single hypothesis: there was a strong mutational AT pressure in the genome of common predecessor of all archaea. Then the rates of AT to GC transitions began to increase, while the rates of AT to GC transversions did not. That is why bias in GC4f and GC2f3p has occurred (GC2f3p>GC4f) and bias in arginine codon usage has not been changed (Arg2>>Arg4) in most of the offspring of common archaeal predecessor. Amino acid content of archaeal proteomes still carries certain features characteristic to proteomes encoded by genomes with strong mutational AT pressure (levels of isoleucine and lysine are increased, while levels of alanine, histidine, cysteine and glutamine are decreased). Many features of archaeal genes and proteins as well as the absence of some genes existing in bacteria and eukaria (*19*) can be explained (at least partially) by our hypothesis.

Materials and Methods

Data

As the material for our *in silico* work we have used 36 lists of codon usage for each coding region (CDS) from 36 completely sequenced archaeal genomes (Tables 1 and 2). All these lists of codon usage for each CDS were taken from Codon Usage Database (www.kazusa.or.jp/codon) (*1*).

Calculation

For the calculation of all necessary indexes needed for the current work, we used “Coding Genome Scanner” (CGS), which is a Microsoft Excel tool containing original algorithm (www.barkovsky.hotmail.ru). The function of CGS is in the calculation of G+C, 1GC, 2GC, 3GC, GC4f, GC2f3p, frequencies of nucleotide, codon and amino acid usage and the entropy of amino acid content distribution for each coding region in the genome.

Coefficients of intragenomic correlation of 3GC on total GC content (R) have been calculated for all species. The coefficient of correlation indicates the strength and direction of a linear relationship between two random variables. The correlation is average or strong if R>0.5 or R<–0.5; the correlation is low or there is no correlation if –0.5<R<0.5.

Entropy of amino acid content distribution (*9*) has been calculated using the Equation 1 from Claude Shannon’s information theory:

$$H = -\sum faa \cdot \log_2 faa \qquad (1)$$

In this equation “faa” is the frequency of amino acid residue usage. The maximum level of uncertainty (H max) for amino acid content of protein is 4,322 bit.

In the second step of our study we mixed all coding regions from the main group of archaea (excluding *H. walsbyi*) and arranged genes according to their GC content (0.2<G+C<0.3; 0.3<G+C<0.4; 0.4<G+C<0.5; 0.5<G+C<0.6; 0.6<G+C<0.7; 0.7<G+C<0.8). Then we compared entropy of amino acid content distribution, average level of 10 a.a. and levels of every amino acid usage in proteins coded by genes from these separate groups using parametric statistics.

The same kind of calculations has been performed on genes and proteins from the group of methane-producing archaea. Then we compared levels of entropy and amino acid levels in groups of proteins coded by genes with the same GC content from the main group of archaea and from the group of methane-producing archaea.

Authors’ contributions

Both authors collected the datasets, conducted data analyses, and co-wrote the manuscript. Both authors read and approved the final manuscript.

Competing interests

The authors have declared that no competing interests exist.

References

1 Nakamura, Y., *et al.* 2000. Codon usage tabulated from the international DNA sequence databases: status for the year 2000. *Nucleic Acids Res.* 28: 292.

2 Sueoka, N. 1988. Directional mutation pressure and neutral molecular evolution. *Proc. Natl. Acad. Sci. USA* 85: 2653-2657.

3 Sueoka, N. 2002. Wide intra-genomic G+C heterogeneity in human and chicken is mainly due to strand-symmetric directional mutation pressures: dGTP-oxidation and symmetric cytosine-deamination hypotheses. *Gene* 300: 141-154.

4 Khrustalev, V.V. and Barkovsky, E.V. 2008. An in-silico study of alphaherpesviruses ICP0 genes: positive selection or strong mutational GC-pressure? *IUBMB Life* 60: 456-460.

5 Gros, L., *et al.* 2002. Enzymology of the repair of free radicals-induced DNA damage. *Oncogene* 21: 8905-8925.

6 Yoon, J.H., *et al.* 2003. Human thymine DNA glycosylase (TDG) and methyl-CpG-binding protein 4 (MBD4) excise thymine glycol (Tg) from a Tg:G mispair. *Nucleic Acids Res.* 31: 5399-5404.

7 Moe, A., *et al.* 2003. Incision at hypoxanthine residues in DNA by a mammalian homologue of the *Escherichia coli* antimutator enzyme endonuclease V. *Nucleic Acids Res.* 31: 3893-3900.

8 Khrustalev, V.V. and Barkovsky, E.V. 2009. Mutational pressure is a cause of inter- and intragenomic differences in GC-content of simplex and varicello viruses. *Comput. Biol. Chem.* 33: 295-302.

9 Khrustalev, V.V. and Barkovsky, E.V. 2009. Main pathways of proteome simplification in alphaherpesviruses under the influence of the strong mutational GC-pressure. *J. Proteomics Bioinform.* 2: 88-96.

10 Fukuchi, S., *et al.* 2003. Unique amino acid composition of proteins in halophilic bacteria. *J. Mol. Biol.* 327: 347-357.

11 Saleh, M., *et al.* 2008. Indicators from archaeal secretomes. *Microbiol Res.* 165: 1-10.

12 Bolhuis, H., *et al.* 2006. The genome of the square archaeon *Haloquadratum walsbyi*: life at the limits of water activity. *BMC Genomics* 7: 169.

13 Falb, M., *et al.* 2008. Metabolism of halophilic archaea. *Extremophiles* 12: 177-196.

14 Cuadros-Orellana, S., *et al.* 2007. Genomic plasticity in prokaryotes: the case of the square haloarchaeon. *ISME J.* 1: 235-245.

15 Khrustalev, V.V. and Barkovsky, E.V. 2007. Levels of CpG and GpC dinucleotides in coding districts of archaeal genomes. In *Proceedings of the Conference on Computational Phylogenetics and Molecular Systematics 2007*, pp.354-357. KMK Scientific Press, Moscow, Russia.

16 Maeder, D.L., *et al.* 2006. The *Methanosarcina barkeri* genome: comparative analysis with *Methanosarcina acetivorans* and *Methanosarcina mazei* reveals extensive rearrangement within methanosarcinal genomes. *J. Bacteriol.* 188: 7922-7931.

17 Galagan, J.E., *et al.* 2002. The genome of *M. acetivorans* reveals extensive metabolic and physiological diversity. *Genome Res.* 12: 532-542.

18 Khrustalev, V.V. and Barkovsky, E.V. 2009. Common predecessor's effect in archaeal genomes and proteomes. In *Proceedings of the Fourth Moscow Conference on Computational Molecular Biology*, pp.163-164. Moscow, Russia.

19 Brown, J.R. and Doolittle, W.F. 1997. Archaea and the prokaryote-to-eukaryote transition. *Microbiol. Mol. Biol. Rev.* 61: 456-502.