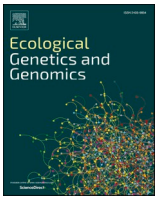
Ecological Genetics and Genomics 25 (2022) 100145

Contents lists available at ScienceDirect

Ecological Genetics and Genomics

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



Chemical evolution of nucleic acids to sustain a life of Archaea

P. Chellapandi *, R. Prathiviraj

Industrial Systems Biology Lab, Department of Bioinformatics, School of Life Sciences, Bharathidasan University, Tiruchirappalli, 620024, Tamil Nadu, India

ARTICLE INFO

Keywords:
Chemical evolution
Cell survivability
Archaea
Growth maintenance
Natural selection
Nucleotides

ABSTRACT

Archaea are a primary domain of the living kingdom, and they play an important role in biogeochemical cycles. Since the inclusion of new archaeal phylogenetic lineages in the universal tree, the origin and evolution history of this domain has been debated. To address this issue, we planned to examine the growth-associated maintenance energy and the proportion of nucleic acids in cell dry weight from 188 archaeal genomes. It was discovered that nucleotide molar fractions influenced evolutionary transmittance across archaeal phyla. At high concentrations of nucleotide molar fractions, minimal cell survivability of archaea was increased. Archaea's survival fitness may have evolved by chemically optimizing the growth-associated maintenance energy required for nucleic acid polymerization. The chemical composition of macromolecules in an archaeal cell may have also acted as a neutral selective pressure shaping its genome dynamics and cell survivability in transient environments. The current hypothesis provides a new look at reduced growth fitness of archaea in a diverse range of environmental niches.

1. Introduction

The modern theory of chemical evolution assumes that on a primitive Earth, a mixture of simple chemicals assembled into more complex molecular systems, from which the first functioning cell emerged [1]. A general concept of chemical evolution is the formation of enantio-enriched biomolecules and the polymerization of simple monomers into information-rich networks. L-amino acids and D-sugars are the fundamental building blocks for two of the most important biological polymer networks (proteins and nucleic acids) required for all forms of life. Within this structural motive, the genetic code is the primary source of all information required for life to exist [2]. The synthesis of life requires the functional integration of various subsystems such as self-replication, metabolism, and compartmentalization that are deemed essential to life. Integrating these characteristics into a single system and allowing it to go through Darwinian evolution should result in the emergence of life [3].

The living system is linked to the initial synthesis and evolution of nucleic acids, which determine a cell's life. The chemical evolution process investigates the concentration of monomers and biomass reactions that allow a cell to survive and thrive on Earth [4]. Nucleic acids are the only molecules capable of coding and transmitting genetic information from generation to generation. DNA and RNA are types of nucleic acids composed of monomers called nucleotides. Nucleotides are not only required for the polymerization of nucleic acids, but they also serve as universal energy transducers in specific cellular functions. A cell's growth rate is related to molar fractions of nucleotides that assort magnificently in various types of cells and cellular systems. However, the proportion of nucleotides and nucleic acids in cell dry weight varies greatly between organisms [5].

Archaea are classified into nine phyla: Euryarchaeota, Crenarchaeota, Thaumarchaeota, Nanoarchaeota, Nanohaloarchaeota, Korarchaeota, Bathyarchaeota, Lokiarchaeota, and Unclassified archaea [6,7]. The availability of a large number of archaeal genomes revealed numerous new insights into the evolution and diversity of these organisms. The archaeal genomes are circular DNAs and range in size from 0.5 to 5.8 Mbp. Several archaeal genomes are made up of multiple chromosomes, each replicated from multiple origins. The archaeal genome is found in a symbiont that derives nutrients from a host, and its small size (<1 Mbp) reflects the deletion of unnecessary genes. The base compositions of archaeal genomes vary greatly, ranging from 28 to 66 mol.% G + C [8]. Phylogenies based on genomes show the pattern of descent among a group of archaeal species. According to Ref. [9]; the first Archaea were anaerobic autotrophs that evolved on the early Earth. Eubacteria and archaea had both evolved independently from the universal ancestor of life (progenote) [10].

Extremophilic bacteria and archaea use a variety of strategies to survive in extreme environments. Increasing the copy number of their

* Corresponding author.
E-mail address: pchellapandi@gmail.com (P. Chellapandi).

https://doi.org/10.1016/j.egg.2022.100145
Received 13 July 2022; Received in revised form 2 September 2022; Accepted 13 September 2022
Available online 22 September 2022

genomes could be one of the adaptive mechanisms of archaea [11]. The ability of archaea to thrive at high temperatures and salinity is a great concern to the scientific community. However, its applicability is limited due to differences in growth physiology and fitness. The molar fractional distribution of nucleotides (mmol monomer/g nucleic acid), growth-associated maintenance (mmol ATP/g nucleic acid), and a proportion of nucleic acid in cell dry weight (mmol nucleic acid/gDCW) were computed as evolutionary constraints to infer the growth fitness of archaea to sustain in extreme environments. These constraints may have shaped archaeal genomes, resulting in adaptation to a new environmental niche [12].

2. Materials and methods

2.1. Dataset

A total of 188 complete genome sequences of archaea were retrieved from the National Collection of Biotechnology Information (www.ncbi.nlm.nih.gov) in FASTA format. The DNA sequences were translated into RNA sequences using BioEdit v7.2 software [13].

2.2. Calculation of nucleotide molar fractions

Metabolic networks are dependent on knowing the chemical composition (nucleic acids, proteins, carbohydrates, and lipids) of the cell and energetic requirements (growth maintenance energy) necessary to generate biomass content from metabolic precursors (nucleotides, amino acids, etc.) [14]. Therefore, the fractional contribution of a nucleotide was estimated from the genome sequences of archaea as described by Ref. [15]. In brief, the molar percentage was multiplied by the molecular weight of the nucleotide to obtain the weight of the nucleotide per mole nucleic acid, which was then added to obtain the weight of the nucleic acid per mole nucleic acid. The weight of nucleotide per mole nucleic acid was converted to weight nucleotide per weight nucleic acid by multiplying by the sum of all nucleotide weights. The weight of a nucleotide was multiplied by the proportion of nucleic acids in a prokaryotic cell [16,17]. This fraction was divided by its molecular weight to obtain the mole nucleotide per cell dry weight. This molar contribution was multiplied by a factor to yield a final unit of mmol nucleotide per gram dry weight. Appendix 1 contains the dataset used to infer nucleic acid chemical evolution.

2.3. Hierarchical cluster analysis

Using a complete linkage method, the calculated values in the dataset were used for hierarchical cluster analysis. Cluster v3.0 software [18] was used to perform the analysis, which generated a dendrogram (CDT format) that was visualized in TreeView v2.0.8 software [19]. The One minus Pearson correlation metric (PCM) and the Euclidian distance metric was used to compute a distance function (EDM). The PCM was also used as a distance measure to determine the linear relationship between genomes [20]. It was calculated by dividing the covariance of the two variables by the product of their standard deviations.

$$PCM = \frac{cov(x,y)}{\sigma x \sigma y}$$

Where *cov* is the convergence, σ_x is the standard deviation of X; σ_y is the standard deviation of y. The EDM was an exhaustive table of distance-square, d_{ij} between points taken by pair from a list of N points in the squared metric, the measure of distance-square [21].

$$EDM = \sqrt{\sum_{i=0}^{k} (x_i - y_i)^2}$$

K is the nearest neighbor measured by a distance function. X_i, is r_{min} (row minimum) and X_j, is r_{max} (row maximum). A heat map of this study was generated by ClusterVis v2.0 [22], Clustergrammer v1.0 [23], and Morpheus tool from Broad Institute (https://software.broadinstitute.org/morpheus/). SYSTAT v13.2 software was used for descriptive, inferential, and variance statistics (Systat Software, Inc.).

3. Results

The current study reconstructed a genome-scale phylogenetic tree of archaea from 188 complete archaeal genome sequences (Fig. 1). The findings of this study show that nine taxonomic classes, as well as two unclassified archaea, such as *Thaumarchaeota, Aigarchaeota, Crenarchaeota*, and *Korarchaeota* (TACK) superphylum (*Proteoarchaeota*) and Deep-Sea Hydrothermal Vent Euryarchaeota 2, are classified separately (DHVE2). TACK superphylum is associated with *Archaeoglobi* and *Nitrososphaeria*, whereas DHVE2 is associated with *Haloarchaea*. Our analysis revealed that mesophilic methanogenic archaea are related to *Haloarchaea*, whereas thermophilic methanogenic archaea are closely related to thermophilic archaea.

The archaeal genome-scale phylogeny was nearly similar to the dendrogram shown in Fig. 2. This dendrogram is divided into four clusters, each of which contains 14 major phylogenetic lineages. An archaeal dendrogram is classified into three main categories based on the proposed evolutionary constraints: low, moderate, and high. Molar fractions increase the phylogeny of archaea at low nucleotide concentrations. It suggests a process of speciation within archaeal phyla at certain concentrations that may have evolved chemically to produce nucleic acids. When nucleotide molar fractions in cells are increased, the conservative nature of archaeal genomes is significantly increased to achieve a stable lineage. A low or moderate quantity of nucleotide molar fractions determines evolutionary transmittance between archaeal phyla. It could be attributed to chemical evolutionary optimization of genome composition and growth-associated nucleic acid polymerization in a cell.

Interestingly, the molar concentration of adenine and thymine increases with decreasing guanine and cytosine concentrations, indicating a low requirement for growth-associated maintenance during DNA synthesis. The proportion of nucleic acid in cell dry weight is not distributed evenly across archaea. During RNA synthesis, the concentrations of adenine and cytosine increase as the concentrations of guanine and uracil decrease. In genome dynamics, growth-associated maintenance and the proportion of nucleic acid in cell dry weight are chemically unbiased. Under this chemical environment, the evolutionary forces acting on archaeal genomes drive them to diversify their genomes, resulting in new species or bifurcation. When the amount of guanine, cytosine, uracil, and ATP required for nucleic acid synthesis increases significantly, the concentrations of adenine and thiamine in a DNA molecule, adenine and cytosine in an RNA molecule, and the proportion of nucleic acids in cell dry weight decrease in archaeal genomes.

4. Discussion

Archaeal lineages are of a major ecological role in modern-day biogeochemical cycles but the genome biology of archaea is not yet completely understood. Genome architecture is conserved between bacteria and archaea. Although Archaea appear to be as old as bacteria, their current diversity is much lower [9]. The near-linear relationship between genome size and the number of encoded proteins may reflect efficient selection against the accumulation of nonfunctional DNA in archaea [24]. The current method inferred the evolutionary imprints of unclassified archaea implicitly. The TACK superphylum is related to the Asgard or Asgardarchaeota superphylum. In the phylogenetic tree, it is affiliated with eukaryotic cellular origins [10,25]. Some TACK and DHVE2 species were found to be closely related to *Thermococci*, *Archaeoglobi*, and *Thaumarchaeota*. However, the lack of antiquity of

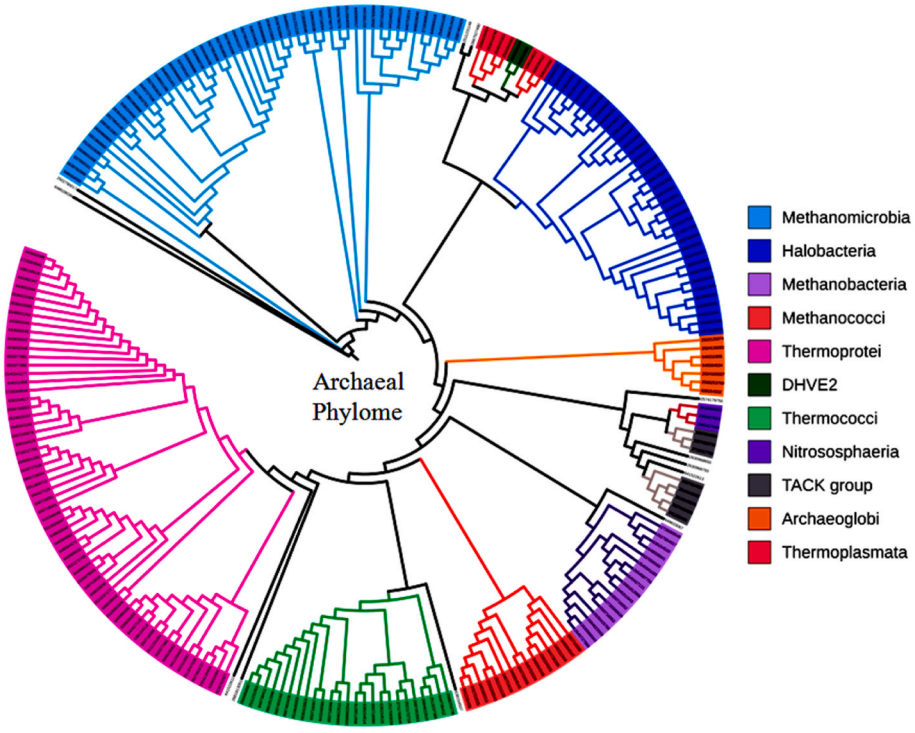


Fig. 1. Ancestral states are deduced from a genome-scale archaeal phylogeny based on COGs in 188 complete genome sequences. A phylogenomic tool included in the IMG/M v5.0 software was used to reconstruct the archaeal phylogenetic lineage (Chen et al., 2017). The phylogenetic distance between genomes was reflected in the average nucleotide identity and alignment fraction values. NSimScan (Novichkov et al., 2016) was used to scan archaeal genomes for nucleotide similarity in the cluster of orthologous groups, which was then filtered to retain bidirectional best hits with at least 70% sequence identity. *Nanoarchaeum equitans* is a tiny hyperthermophilic symbiont of *Ignicoccus hospitalis*, a Crenarchaeote. It served as an outlier organism. The colors of the branches represent archaeal classes at the tree's tips and inferred states at ancestral nodes. (For interpretation of the references to color in this figure legend, the reader is referred to the Web version of this article.)

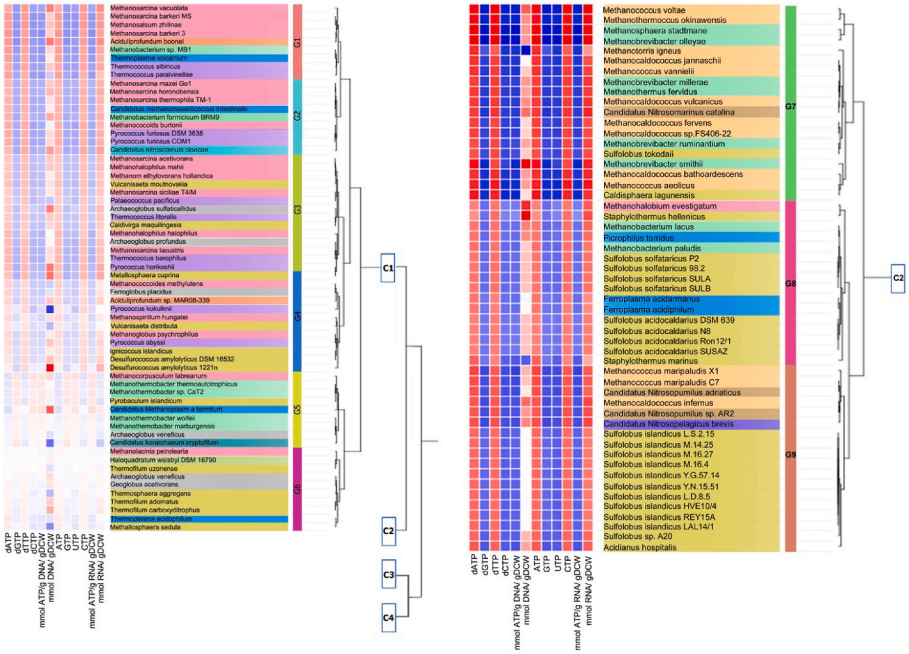


Fig. 2. A dendrogram based on the molar fractional distribution of nucleotides (mmol monomer/g nucleic acid), growth-associated maintenance (mmol ATP/g nucleic acid), and a proportion of nucleic acid in cell dry weight (mmol nucleic acid/gDCW) for inferring the chemical evolution of nucleic acids in archaea. Vertical color bars (G1-G14) represent archaeal lineage sub-clades from clusters 1 to 4. (Cluster 1 is on the top left, Cluster 2 is on the right, Cluster 3 is on the bottom left, and Cluster 3 is on the bottom right). (For interpretation of the references to color in this figure legend, the reader is referred to the Web version of this article.)

archaeal fossil traces may severely affect any attempt to date the origin of this domain by molecular data [26]. Thus, the addition of a robust genome-scale phylogeny to the current approach has provided a good framework for reconstructing the evolutionary history of this domain.

As its macromolecules evolved more slowly, any hyperthermophilic lineage retained many ancestral characteristics [27]. Herewith we revealed a common hyperthermophilic origin for the evolution of the euryarchaeal and crenarchaeal phyla, which is consistent with previous research [28,29]. We also deduced the evolutionary transmittance between TACK and DHVE2; crenarchaeota and euryarchaeota at low nucleotide concentrations. It was found to be in good agreement with previous phylogenetic inferences [30]. It implied that the concentration of nucleotide molar fractions was a determinant of archaeal genome bifurcation into different orders and phyla. Amelioration is a fundamental neutral selective pressure in prokaryotes that shapes intergenic base composition to evolutionary history and environmental adaptations [31,32]. The fraction of free nucleotide positions is an important determinant of DNA divergence over time, and it was used to explain differences in DNA divergence rates [33]. The mutational pressure that leads to nucleotide substitutions is also highly correlated with the DNA composition of archaeal genomes. The proportion of each type of nucleotide in archaeal genomes is proportional to the time required to

replace half of the nucleotides [34].

In archaea, adenine and cytosine are swerving, while thymine and guanine are distorted [35]. Our findings show that increasing adenine and thymine concentrations while decreasing guanine, cytosine, and ATP concentrations are required for DNA synthesis. Guanine and uracil concentrations were directly proportional to adenine and cytosine concentrations in archaea for RNA polymerization. The evolutionary adaptation to oscillated environments determines the rate of minimal growth [36]. Growth-associated maintenance was chemically unbiased with a proportion of nucleic acid in the cell, which could significantly increase the archaea genome dynamic rate. Archaea chemically evolved in a specific cluster or lineage with more growth-associated maintenance energy has lower growth fitness. As a result, the molar concentration of nucleotides and growth-associated maintenance energy is directly proportional to the conservation of archaeal genome composition and reduced growth fitness [37]. However, there is an indirect link to amino acid requirements to make the proteome composition for the survival fitness of archaeal cells.

The evolution of nucleic acids is of great interest for gaining a better understanding of genome replication machinery, lineage- and niche-specific adaptation, codon usage, and amino acid diversity [32]. Several models for studying nucleotide evolution based on substitution rates and frequencies have been developed [38]. A super-statistical model has been developed to investigate non-trivial universality in bacterial DNA architecture inter-nucleotide interval distributions [39]. In this study, the molar fraction of nucleic acids was considered as an evolutionary constraint to infer the growth fitness of archaea, which may have shaped archaeal genomes to specific environmental conditions.

5. Conclusions

Despite fossil traces and the sequence-based tree of life, the nucleotide molar fraction is the first reliable method for studying archaea genome biology and ancestry. Our proposed nucleotide constraints and tenancy in biomass chemical constituents are chemically converged in the archaeal domain. Interestingly, ATP maintenance energy (mmol/g nucleic acid) is inversely proportional to the molar fraction of nucleic acids in biomass composition (mmol/gDCW). These chemical constraints may act as neutral selective forces on archaea genome dynamics, evolution, growth fitness, and cell survivability. The current approach also provides some progress in the taxonomic placement of unclassified archaea (TACK and DHVE2 groups) in modern-day archaeal lineages.

CRediT authorship contribution statement

P. Chellapandi: Concept of the work and manuscript writing and revision. **R. Prathiviraj:** Data collection, interpretation, and analysis.

Declaration of competing interest

The authors confirm that this article's content has no conflicts of interest.

Data availability

Data will be made available on request

Acknowledgments

The authors would like to thank the University Grants Commission (30–1/2013(SA-II)/RA-2012-14-SC-TAM-1768 and 42–864/2013-SR), the Government of India, for financial assistance.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.egg.2022.100145.

References

[1] L.G. Pleasant, C. Ponnamperuma, Chemical evolution and the origin of life: bibliography supplement 1981, Orig. Life 13 (1983) 61–80.
[2] R. Krishnamurthy, N.V. Hud, Introduction: chemical evolution and the origins of life, Chem. Rev. 120 (2020) 4613–4615.
[3] S. Otto, An approach to the *de novo* synthesis of life, Acc. Chem. Res. 55 (2022) 145–155.
[4] S.I. Walker, M.A. Grover, N.V. Hud, Universal sequence replication, reversible polymerization and early functional biopolymers: a model for the initiation of prebiotic sequence evolution, PLoS One 7 (2012) 1–12.
[5] B. Alberts, A. Johnson, J. Lewis, M. Raff, K. Roberts, P. Walter, in: Molecular Biology of the Cell, 4th edition, Garland Science, New York, 2002.
[6] A. Spang, E.F. Caceres, T.J.G. Ettema, Genomic exploration of the diversity, ecology, and evolution of the archaeal domain of life, Science 357 (2017), eaaf3883.
[7] A. Spang, T.J.G. Ettema, Archaeal evolution: the methanogenic roots of Archaea, Nat. Microbiol. 2 (2017), 17109.
[8] P. Chellapandi, Molecular evolution of methanogens based on their metabolic facets, Front. Biol. 6 (2011) 490–503.
[9] S. Kellner, A. Spang, P. Offre, G.J. Szöllősi, C. Petitjean, T.A. Williams, Genome size evolution in the Archaea, Emerg Top Life Sci 2 (2018) 595–605.
[10] S.J. Berkemer, S.E. McGlynn, A new analysis of archaea-bacteria domain separation: variable phylogenetic distance and the tempo of early evolution, Mol. Biol. Evol. 37 (2020) 2332–2340.
[11] K.A. Dulmage, C.L. Darnell, A. Vreugdenhil, A.K. Schmid, Copy number variation is associated with gene expression change in archaea, Microb. Genom. 4 (2018), e000210.
[12] K.U. Foerstner, C. von Mering, S.D. Hooper, P. Bork, Environments shape the nucleotide composition of genomes, EMBO Rep. 6 (2005) 1208–1213.
[13] T.A. Hall, BioEdit: a user-friendly biological sequence alignment editor and analysis program for windows 95/98/NT, Nucleic Acids Symp. Ser. 41 (1999) 95–98.
[14] J.C. Lachance, C.J. Lloyd, J.M. Monk, L. Yang, A.V. Sastry, Y. Seif, B.O. Palsson, S. Rodrigue, A.M. Feist, Z.A. King, P.É. Jacques, BOFdat: generating biomass objective functions for genome-scale metabolic models from experimental data, PLoS Comput. Biol. 15 (2019), e1006971.
[15] I. Thiele, B.Ø. Palsson, A protocol for generating a high-quality genome-scale metabolic reconstruction, Nat. Protoc. 5 (2010) 93–121.
[16] F.C. Neidhardt, J.L. Ingraham, M. Schaechter, Physiology of the Bacterial Cell. A Molecular Approach, Sinauer associates, Sunderland, MA, 1990, p. 507.
[17] M.N. Benedict, M.C. Gonnerman, W.W. Metcalf, N.D. Price, Genome-scale metabolic reconstruction and hypothesis testing in the methanogenic archaeon *Methanosarcina acetivorans* C2A, J. Bacteriol. 194 (2012) 855–865.
[18] M.J. de Hoon, S. Imoto, J. Nolan, S. Miyano, Open source clustering software, Bioinformatics 20 (2004) 1453–1454.
[19] R.D. Page, Visualizing phylogenetic trees using TreeView, Curr Protoc Bioinformatics (2002) (Chapter 6):Unit 6.2.
[20] J.G. Wagner, G.K. Aghajanian, O.H. Bing, Correlation of performance test scores with tissue concentration of lysergic acid diethylamide in human subjects, Clin. Pharmacol. Therapeut. 9 (1968) 635–638.
[21] Dattorro, Convex optimization, euclidean distance geometry 2ε, Mεβoo (2015) v2015.07.21.
[22] T. Metsalu, J. Vilo, ClustVis: a web tool for visualizing clustering of multivariate data using Principal Component Analysis and heatmap, Nucleic Acids Res. 43 (2015) W566–W570.
[23] N.F. Fernandez, G.W. Gundersen, A. Rahman, M.L. Grimes, K. Rikova, P. Hornbeck, A. Ma'ayan, Clustergrammer, a web-based heatmap visualization and analysis tool for high-dimensional biological data, Sci. Data 4 (2017), 170151.
[24] M. Lynch, Streamlining and simplification of microbial genome architecture, Annu. Rev. Microbiol. 60 (2006) 327–349.
[25] Q. Zhu, U. Mai, W. Pfeiffer, S. Janssen, F. Asnicar, J.G. Sanders, P. Belda-Ferre, G. A. Al-Ghalith, E. Kopylova, D. McDonald, T. Kosciolek, J.B. Yin, S. Huang, N. Salam, J.Y. Jiao, Z. Wu, Z.Z. Xu, K. Cantrell, Y. Yang, E. Sayyari, M. Rabiee, J. T. Morton, S. Podell, D. Knights, W.J. Li, C. Huttenhower, N. Segata, L. Smarr, S. Mirarab, R. Knight, Phylogenomics of 10,575 genomes reveals evolutionary proximity between domains Bacteria and Archaea, Nat. Commun. 10 (2019) 5477.
[26] T. Cavalier-Smith, The neomuran origin of archaebacteria, the negibacterial root of the universal tree and bacterial megaclassification, Int. J. Syst. Evol. Microbiol. 52 (2002) 7–76.
[27] O. Matte-Tailliez, C. Brochier, P. Forterre, H. Philippe, Archaeal phylogeny based on ribosomal proteins, Mol. Biol. Evol. 19 (2002) 631–639.
[28] P. Forterre, A hot story from comparative genomics: reverse gyrase is the only hyperthermophile-specific protein, Trends Genet. 18 (2002) 236–237.
[29] S. Gribaldo, C. Brochier-Armanet, The origin and evolution of Archaea: a state of the art, Philos. Trans. R. Soc. Lond. B Biol. Sci. 361 (2006) 1007–1022.
[30] M. Bharathi, P. Chellapandi, Intergenomic evolution and metabolic cross-talk between rumen and thermophilic autotrophic methanogenic archaea, Mol. Phylogenet. Evol. 107 (2017) 293.

[31] S. Mann, Y.P. Chen, Bacterial genomic G+C composition-eliciting environmental adaptation, Genomics 95 (2010) 7–15.
[32] J. Bohlin, V. Eldholm, J.H. Pettersson, O. Brynildsrud, L. Snipen, The nucleotide composition of microbial genomes indicates differential patterns of selection on core and accessory genomes, BMC Genom. 18 (2017) 151.
[33] S.R. Palumbi, Rates of molecular evolution and the fraction of nucleotide positions free to vary, J. Mol. Evol. 29 (1989) 180–187.
[34] M. Kowalczuk, P. Mackiewicz, D. Mackiewicz, A. Nowicka, M. Dudkiewicz, M. R. Dudek, S. Cebrat, High correlation between the turnover of nucleotides under mutational pressure and the DNA composition, BMC Evol. Biol. 1 (2001) 13.
[35] R.S. Gupta, Protein phylogenies and signature sequences: a reappraisal of evolutionary relationships among archaebacteria, eubacteria, and eukaryotes, Microbiol. Mol. Biol. Rev. 62 (1998) 1435–1491.
[36] B.W. Ying, T. Honda, S. Tsuru, S. Seno, H. Matsuda, Y. Kazuta, T. Yomo, Evolutionary consequence of a trade-off between growth and maintenance along with ribosomal damages, PLoS One 10 (2015), e0135639.
[37] M. Pagel, Inferring the historical patterns of biological evolution, Nature 401 (1999) 877–884.
[38] W.H. Li, Molecular Evolution, Sinauer Associates, Inc., Sunderland MA, 1997.
[39] M.I. Bogachev, O.A. Markelov, A.R. Kayumov, A. Bunde, Superstatistical model of bacterial DNA architecture, Sci. Rep. 7 (2017), 43034.