Genome DNA Sequence Variation, Evolution, and Function in Bacteria and Archaea

Hiromi Nishida*

Agricultural Bioinformatics Research Unit, Graduate School of Agricultural and Life Sciences, University of Tokyo, Japan

Abstract
Comparative genomics has revealed that variations in bacterial and archaeal genome DNA sequences cannot be explained by only neutral mutations. Virus resistance and plasmid distribution systems have resulted in changes in bacterial and archaeal genome sequences during evolution. The restriction-modification system, a virus resistance system, leads to avoidance of palindromic DNA sequences in genomes. Clustered, regularly interspaced, short palindromic repeats (CRISPRs) found in genomes represent yet another virus resistance system. Comparative genomics has shown that bacteria and archaea have failed to gain any DNA with GC content higher than the GC content of their chromosomes. Thus, horizontally transferred DNA regions have lower GC content than the host chromosomal DNA does. Some nucleoid-associated proteins bind DNA regions with low GC content and inhibit the expression of genes contained in those regions. This form of gene repression is another type of virus resistance system. On the other hand, bacteria and archaea have used plasmids to gain additional genes. Virus resistance systems influence plasmid distribution. Interestingly, the restriction-modification system and nucleoid-associated protein genes have been distributed via plasmids. Thus, GC content and genomic signatures do not reflect bacterial and archaeal evolutionary relationships.

Distribution of genome base compositions and mutational biases
Among all published genome sequences, Candidatus Zinderia insecticola has a genome with the lowest guanine-cytosine (GC) content (13.5%) (McCutcheon and Moran, 2010), and Anaeromyxobacter dehalogenans 2CP-C has a genome with the highest GC content (74.9%) (Thomas et al., 2008). The distribution of GC content of bacterial genomes is rather distinct from a normal (Gaussian) distribution (Figure 1). On the other hand, within each bacterium, the distribution of the GC content of the genes is similar to a normal distribution (Figure 2), suggesting that each bacterium has maintained its genomic GC content. Bacteria have been thought to possess directionality, driven by neutral forces, toward higher or lower levels of GC content.

*Corresponding author: ahn@mail.ecc.u-tokyo.ac.jp

Figure 1. Distribution of GC content of bacterial genomes.
content in their DNA (Sueoka, 1961; Freese, 1962; Sueoka, 1962; Sueoka, 1988).

However, recent studies have shown that the variation of the GC content among bacteria is driven by selection, in which mutations from GC to adenosine-thymine (AT) are more common than mutations from AT to GC (Hershberg and Petrov, 2010; Hildebrand et al., 2010; Rocha and Feil, 2010). Lind and Andersson (2008) compared the genomes of 2 Salmonella typhimurium mutants and showed a bias toward mutations from GC to AT. Rocha and Danchin (2002) suggested that GC content variation may be related to the higher energy cost and limited availability of G and C over A and T. However, many bacterial species such as Actinobacteria have a high GC content genome. How have these bacteria maintained a high GC content? DNA polymerase components involved in DNA replication have been reported to directly influence the GC contents of genomes (Zhao et al., 2007; Wu et al., 2012).

**GC content and genomic signature**

Oligonucleotide frequencies (genomic signatures) within a genome can be observed and compared with other genomes (Campbell et al., 1999; Deschavanne et al., 1999). Undoubtedly, genomes with similar genomic signatures have similar GC contents. Interestingly, genomes with similar GC contents also have similar genomic signatures (Albrecht-Buehler, 2007a; Albrecht-Buehler, 2007b; Zhang and Wang, 2011), with the exception of Deinococcus radiodurans and Thermus thermophilus (Nishida et al., 2012a). Phylogenetic relationships based on genomic signature comparison of 89 bacteria (Nishida et al., 2012a) were found to be completely different from those based on gene content or orthologous protein sequence comparison (Nishida et al., 2011). This indicates that organisms with genomic signature similarity do not represent closely related organisms in evolutionary terms (Albrecht-Buehler, 2007a; Bohlin, 2011).

---

**Figure 2.** Distribution of GC content of genes of three bacteria species, *Mycoplasma mycoides* (genome with low GC content), *Escherichia coli* (genome with middle GC content), and *Streptomyces griseus* (genome with high GC content).
In addition, frequencies of palindromic DNA sequence patterns are significantly lower than those of non-palindromic sequence patterns in bacterial and archaeal genomes (Gelfand and Koonin, 1997). Palindrome avoidance has been reported to be intimately correlated with infective behavior of the bacteriophage (Rocha et al., 2001). The low frequency of palindromic sequence patterns has been found in not only single genome sequence but also metagenomic sequence data (Dick et al., 2009). Generally, restriction enzymes recognize palindromic DNA sequences and digest these regions. However, when palindromic DNA sequences are methylated by a DNA methylase, restriction enzymes can no longer digest them (Wilson and Murray, 1991; Bickle and Krüger, 1993). Bacteria protect their palindromic DNAs modified by modification enzymes, but digest bacteriophage palindromic DNAs that are not modified. Thus, the restriction-modification system functions as a virus resistance system (Kobayashi, 2001; Labrie et al., 2010). Palindrome avoidance influences genomic signatures.

Surprisingly, bacteria and archaea possess clustered, regularly interspaced, short palindromic repeats (CRISPRs) and the CRISPR-associated (cas) genes as a virus resistance system, which acts as a defense system against viral infections through the use of CRISPR RNA transcripts (Barrangou et al., 2007; Brouns et al., 2008; Marraffini and Sontheimer, 2008; Sorek et al., 2008; Karginov and Hannon, 2010; Labrie et al., 2010). Thus, bacteria and archaea employ palindromic DNA sequence avoidance as well as palindromic DNA sequences in CRISPR regions, as a defense system against viral infections.

Restriction-modification systems may also influence GC content and genomic signatures. DNA methylase distributions among bacteria and archaea were shown in Figure 3. *Escherichia coli* has DNA adenine methylase gene (*dam*; NCBI gene ID, 947893) and DNA cytosine methylase gene (*dcm*; NCBI gene ID, 946479) (Marinus, 1987). Based on the Kyoto Encyclopedia of Genes and Genomes (KEGG) orthology database (Kanehisa et al., 2012), the distributions of *dam* and *dcm* differ (Figure 3). The differences in their distributions suggest that genomic GC content is related to DNA methylase distribution. For example, most Actinobacteria (genomes with high GC content) lack *dam*, and most Spirochaetes (genomes with low GC content) lack *dcm*. However, certain cases cannot be explained by genome GC content. Most Tenericutes (genomes with low GC content) have both *dam* and *dcm* and lack *dcm*. In addition, although the GC contents of Crenarchaeota are not high (35–60% among 40 archaea used in this study), all Crenarchaeota possess *dcm* and lack *dam*. On the other hand, restriction enzymes have no structural similarity with DNA methylases or other restriction enzymes (Bickle and Krüger, 1993). It was reported that the restriction-modification system is a mobile element (Kusano et al., 1995; Naito et al., 1995; Kobayashi, 2001).
Genomic GC content influences codon and amino acid usages (Muto and Osawa, 1987; Lobry, 1997; Singer and Hickey, 2000; Knight et al., 2001; Chen et al., 2004; Wan et al., 2004; Lightfield et al., 2011; Schmidt et al., 2012). In addition, the GC contents of horizontally transferred DNA regions have been ameliorated to adjust to host chromosome GC content (Lawrence and Ochman, 1997). It strongly suggests that each bacterium or archaeon has a system for maintaining organism-specific GC content and genomic signature.

**Differences in GC contents between host genome DNA and horizontally transferred DNA regions**

Genome size and GC content are weakly correlated in bacteria and archaea (Bentley and Parkhill, 2004; Musto et al., 2006; Mitchell, 2007; Suzuki et al., 2008; Guo et al., 2009; Nishida, 2012). The genomes of obligate host-associated bacteria are short and low GC content (Moran, 2002; Klasson and Andersson, 2004; McCutcheon and Moran, 2012), with exception of *Candidatus* Hodgkinia cicadicola (McCutcheon et al., 2009; Van Leuven and McCutcheon, 2012). In addition, horizontally transferred DNA, plasmid DNA, and virus DNA have lower GC content than host chromosome DNAs do (Rocha and Danchin, 2002). Most of the differences in GC content between plasmids and their host chromosomes are of less than 10% (Nishida, 2012), suggesting that host organisms cannot maintain and regulate plasmids with very different GC content from their own. If bacteria and archaea maintain lower GC content for horizontally transferred regions, this maintenance will compete with their amelioration of GC content to match the GC content of the host genome (Lawrence and Ochman, 1997).

Interestingly, bacteria and archaea have not acquired DNAs with a GC content higher than the GC content of their own genome. Bacterial and archaeal genomes with high AT content are protected from attacks by most viruses. On the other hand, it is difficult for those organisms to use any plasmids. The genome sizes of obligate host-associated bacteria are decreasing (for example, Oshima and Nishida, 2008). However, the genome size reduction is not limited in obligate host-associated bacteria (Nilsson et al., 2005). There is a general bias among bacteria toward genomic deletions rather than insertions (Mira et al., 2004; Lightfield et al., 2011; Schmidt et al., 2012). It is possible that obligate host-associated bacteria do not need additional gene uptake. It may therefore be hypothesized that these bacteria maintain a genome with low GC content as a virus resistance system.

**GC content and nucleoid-associated proteins**

Nucleoid-associated proteins are related not only to nucleoid structures but also to gene regulation (Dillon and Dorman, 2010). The heat-stable (or histone-like) nucleoid-structural (H-NS) protein in *Salmonella enterica* binds DNA regions with low GC content rather than the remaining chromosomal DNA and inhibits expression of the genes contained in those regions, to which horizontally transferred DNA fragments locate (Lucchini et al., 2006; Navarre et al., 2006). Similar functions for nucleoid-associated proteins have been found in other bacteria (Castang et al., 2008; Gordon et al., 2010; Smits and Grossman, 2010; Yun et al., 2010). These gene-silencing systems depend on the fact that horizontally transferred DNAs have lower GC content than host chromosome DNAs do (Rocha and Danchin, 2002).

This gene-repression system involving nucleoid-associated proteins is widespread amongst bacteria and archaea, suggesting that the nucleoid-associated proteins may bind to DNA regions with different GC content between different bacterial or archaeal species. For example, in the *Symbiobacterium thermophilus* genome with high GC content (69%), transposase genes, markers of transposable genetic elements, are more frequently found in regions with lower GC content (less than 65% GC content) than in the remaining chromosomal DNA (Nishida and Yun, 2011). Interestingly, nucleoid-associated protein genes are distributed not only throughout bacterial chromosomes but also within plasmids, suggesting that plasmids have carried these genes (Yun et al., 2010; Takeda et al., 2011).

Although nucleoid-associated proteins have different structures, they share the same function (Gordon et al., 2011). Interestingly, core histones, which are structurally different from bacterial nucleoid-associated proteins, prefer AT-rich DNA to GC-rich DNA. This DNA sequence preference plays an important role in nucleosome formation (Segal et al., 2006; Segal and Widom, 2009; Valouev et al., 2011; Nishida et al., 2012b). The interactions between DNA sequence preferences and nucleoid-associated proteins may have played an important role in global regulations of genes among Bacteria, Archaea, and Eukarya during evolution.

**Acknowledgements**

I thank Professor Teruhiko Beppu for his valuable comments and encouragements.

**References**


Further Reading

Caister Academic Press is a leading academic publisher of advanced texts in microbiology, molecular biology and medical research. Full details of all our publications at caister.com

- MALDI-TOF Mass Spectrometry in Microbiology
  Edited by: M Kostrzewa, S Schubert (2016)
  www.caister.com/malditof

- Aspergillus and Penicillium in the Post-genomic Era
  Edited by: RP Vries, IB Gelber, MR Andersen (2016)
  www.caister.com/aspergillus2

- The Bacteriocins: Current Knowledge and Future Prospects
  Edited by: RL Dottt, SM Roy, MA Riley (2016)
  www.caister.com/bacteriocins

- Omics in Plant Disease Resistance
  Edited by: V Bhadauria (2016)
  www.caister.com/opdr

- Acidophiles: Life in Extremely Acidic Environments
  Edited by: R Quatrini, DB Johnson (2016)
  www.caister.com/acidophiles

- Climate Change and Microbial Ecology: Current Research and Future Trends
  Edited by: J Marxsen (2016)
  www.caister.com/climate

- Biofilms in Bioremediation: Current Research and Emerging Technologies
  Edited by: G Lear (2016)
  www.caister.com/biorem

- Microalgae: Current Research and Applications
  Edited by: MN Tsaioglou (2016)
  www.caister.com/microalgae

- Gas Plasma Sterilization in Microbiology: Theory, Applications, Pitfalls and New Perspectives
  Edited by: H Shintani, A Sakudo (2016)
  www.caister.com/gasplasma

- Virus Evolution: Current Research and Future Directions
  Edited by: SC Weaver, M Denison, M Roossinck, et al. (2016)
  www.caister.com/virusevol

- Arboviruses: Molecular Biology, Evolution and Control
  Edited by: N Vasilakis, DJ Gubler (2016)
  www.caister.com/arbo

- Shigella: Molecular and Cellular Biology
  Edited by: WD Picking, WL Picking (2016)
  www.caister.com/shigella

- Aquatic Biofilms: Ecology, Water Quality and Wastewater Treatment
  Edited by: AM Romani, H Guasco, MD Balaguer (2016)
  www.caister.com/aquaticbiofilms

- Alphaviruses: Current Biology
  Edited by: S Mahalingam, L Herrero, B Herring (2016)
  www.caister.com/alpha

- Thermophilic Microorganisms
  Edited by: F Li (2015)
  www.caister.com/thermophile

- Flow Cytometry in Microbiology: Technology and Applications
  Edited by: MG Wilkinson (2015)
  www.caister.com/flow

- Probiotics and Prebiotics: Current Research and Future Trends
  Edited by: K Venema, AP Carmon (2015)
  www.caister.com/probiotics

- Epigenetics: Current Research and Emerging Trends
  Edited by: BP Chadwick (2015)
  www.caister.com/epigenetics2015

- Corynebacterium glutamicum: From Systems Biology to Biotechnological Applications
  Edited by: A Burkovski (2015)
  www.caister.com/cory2

- Advanced Vaccine Research Methods for the Decade of Vaccines
  Edited by: F Bagnoli, R Rappuoli (2015)
  www.caister.com/vaccines

- Antifungals: From Genomics to Resistance and the Development of Novel Agents
  Edited by: AT Coste, P Vandeputte (2015)
  www.caister.com/antifungals

- Bacteria-Plant Interactions: Advanced Research and Future Trends
  www.caister.com/bacteria-plant

- Aeromonas
  Edited by: J Graf (2015)
  www.caister.com/aeromonas

- Antibiotics: Current Innovations and Future Trends
  Edited by: S Sánchez, AL Demain (2015)
  www.caister.com/antibiotics

- Leishmania: Current Biology and Control
  Edited by: S Adak, R Datta (2015)
  www.caister.com/leish2

- Acanthamoeba: Biology and Pathogenesis (2nd edition)
  Author: NA Khan (2015)
  www.caister.com/acanthamoeba2

- Microarrays: Current Technology, Innovations and Applications
  Edited by: Z He (2014)
  www.caister.com/microarrays2

- Metagenomics of the Microbial Nitrogen Cycle: Theory, Methods and Applications
  Edited by: D Marco (2014)
  www.caister.com/n2

Order from caister.com/order