SAGE-Hindawi Access to Research
International Journal of Evolutionary Biology
Volume 2011, Article ID 634505, 5 pages
doi:10.4061/2011/634505

Research Article

Phylogenetic and Guanine-Cytosine Content Analysis of *Symbiobacterium thermophilum* Genes

Hiromi Nishida and Choong-Soo Yun

Agricultural Bioinformatics Research Unit, Graduate School of Agriculture and Life Sciences, The University of Tokyo, Bunkyo-ku, Tokyo 113-8657, Japan

Correspondence should be addressed to Hiromi Nishida, hnishida@iu.a.u-tokyo.ac.jp

Received 10 September 2010; Revised 20 October 2010; Accepted 5 November 2010

Academic Editor: Shinji Kondo

Although the bacterium *Symbiobacterium thermophilum* has a genome with a high guanine-cytosine (GC) content (69%), it belongs to a low GC content bacterial group. We detected only 18 low GC content regions with 5 or more consecutive genes whose GC contents were below 65% in the genome of this organism. *S. thermophilum* has 66 transposase genes, which are markers of transposable genetic elements, and 38 (58%) of them were located in the low GC content regions, suggesting that *Symbiobacterium* has a similar gene silencing system as *Salmonella*. The top hit (best match) analyses for each *Symbiobacterium* protein showed that putative horizontally transferred genes and vertically inherited genes are scattered across the genome. Approximately 25% of the 3338 *Symbiobacterium* proteins have the highest similarity with the protein of a phylogenetically distant organism. The putative horizontally transferred genes also have a high GC content, suggesting that *Symbiobacterium* has gained many DNA fragments from phylogenetically distant organisms during the early stage of Firmicutes evolution. After acquiring genes, *Symbiobacterium* increased the GC content of the horizontally transferred genes and thereby maintained a genome with a high GC content.

1. Introduction

Symbiobacterium thermophilum is a syntrophic bacterium that grows effectively when cocultured with a cognate *Geobacillus* sp. [1]. Because of the lack of carbonic anhydrase in the course of *Symbiobacterium* evolution [2, 3], the major growth factor for this organism is CO_2 generated by the growth of *Geobacillus* [4]. *S. thermophilum* has a 3.57 Mbp circular genome that consists of 3338 protein-coding sequences [3]. On the basis of the comparative genomic studies, *Symbiobacterium* is classified as a member of the class Clostridia [3, 5]. Although *Symbiobacterium* phylogenetically belongs to Clostridia (low guanine-cytosine (GC) content bacterial group), the species *S. thermophilum* has a genome with a high GC content (69%).

GC content is commonly used as a marker in bacterial systematics; for example, actinobacteria have a high GC content genome, and clostridia have a low GC content genome. This variation in nucleotide content in bacteria is not clearly understood [6–8]. Analyzing a high GC content genome of a bacterium that belongs to a low GC content group or vice versa is useful and important. *Symbiobacterium* belongs to the class Clostridia (low GC content group), but its genome has a high GC content (69%). One possibility is that *Symbiobacterium* has acquired this high GC content from DNA fragments through horizontal gene transfer [9, 10], homologous gene recombination [11, 12], or both. Another possibility is that *Symbiobacterium* has increased the GC content of the acquired genes and maintained the high GC content during evolution. In this study, we identified GC content of each gene of *S. thermophilum*. In addition, we identified the horizontally transferred and vertically inherited genes. In order to elucidate why the *Symbiobacterium* genome has a high GC content, we compared the GC contents of the horizontally transferred genes with those of the vertically inherited genes.

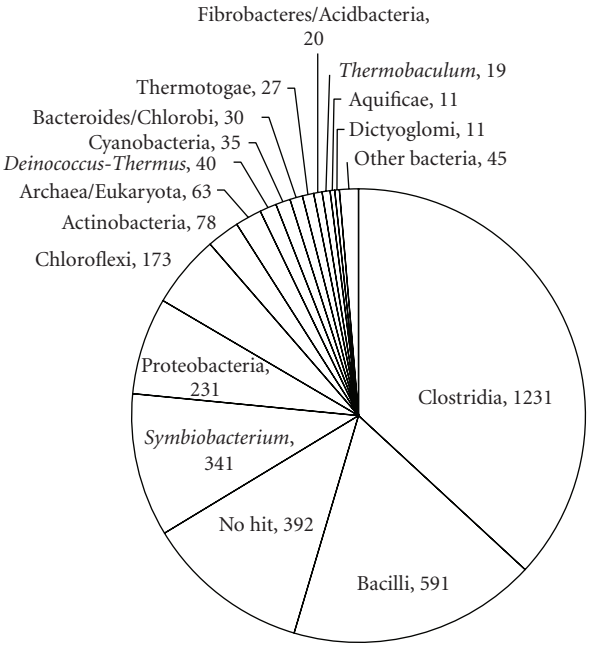

Figure 1: Pie chart of the categories of the 3338 *Symbiobacterium* protein-coding genes. A BLAST search was conducted for all proteins from 147 eukaryotes, 1047 bacteria, and 84 archaea in the KEGG database (http://www.kegg.jp) considering the parameter values given on the GenomeNet website (http://www.genome.jp). The query amino acid sequence was each protein of *Symbiobacterium thermophilum*. The top hit (best match) for each *Symbiobacterium* protein was recorded. However, if the top hit was absent or if the *E*-value of the top hit exceeded 0.1, the *Symbiobacterium* protein was considered to have no similar protein (category, "No hit"). We categorized the 3338 *Symbiobacterium* proteins into the following 17 categories: "Actinobacteria," "Aquificae," "Archaea/Eukaryota," "Bacilli," "Bacteroidetes/Chlorobi," "Chloroflexi," "Clostridia," "Cyanobacteria," "*Deinococcus-Thermus,*" "Dictyoglomi," "Fibrobacteres/Acidobacteria," "No hit," "Proteobacteria," "*Symbiobacterium,*" "*Thermobaculum,*" "Thermotogae," and "Other bacteria." If the top hit was another protein(s) of *Symbiobacterium*, then the query protein was considered to belong to the category "*Symbiobacterium.*"

2. Materials and Methods

In this study, we classified the 3338 protein-coding sequences of *S. thermophilum* on the basis of the amino acid sequence of each coded protein. A BLAST search was conducted for all proteins from 147 eukaryotes, 1047 bacteria, and 84 archaea in the KEGG database (http://www.kegg.jp/) considering the parameter values given on the GenomeNet website (http://www.genome.jp/). The top hit (best match) for each *Symbiobacterium* protein was recorded. However, if the top hit was missing or if the *E*-value of the top hit exceeded 0.1, we considered the *Symbiobacterium* protein to have no similar protein (category, "No hit"). If the top hit was another protein(s) of *Symbiobacterium* (category, "*Symbiobacterium*"), the protein-coding gene was considered to have duplicated during evolution. Top hit analysis at the genome level is a powerful tool for elucidating the phylogenetic lineage of an organism [13, 14].

3. Results and Discussion

On the basis of the phylogenetic lineage of the organism possessing the top hit protein shown in the BLAST result, the 3338 *S. thermophilum* protein-coding genes were classified into 17 categories (Figure 1). The largest category was "Clostridia," and the second largest category was "Bacilli." This is consistent with the results of previous phylogenetic analyses [3]. The third and fourth largest categories were "No hit" and "*Symbiobacterium,*" respectively (Figure 1). Most genes belonging to the category "*Symbiobacterium*" might share their origin with other genes of the same category because 300 of the 341 genes had a similar protein sequence as that of the other organisms that appeared below the top hit of the BLAST result (Table 1 (see Supplementary matrial available online at doi:10.4061/2011/634505)). For example, most transposable elements belonged to "*Symbiobacterium,*" indicating that they were duplicated on the *Symbiobacterium* genome after invasion.

When each gene was plotted on the basis of its category, we detected 52 clusters containing 5 or more consecutive genes belonging to "Clostridia" (Figure 2, pink regions in Supplementary Table 1). These conserved gene clusters are probably not acquired by horizontal gene transfer and are strongly considered to be vertically inherited. The putative vertically inherited genes were scattered across the genome of *S. thermophilum* (Figure 2). In addition, we detected 18 low GC content regions containing 5 or more consecutive genes whose GC contents were below 65% (Figure 3, yellow regions in Supplementary Table 1). These low GC content regions

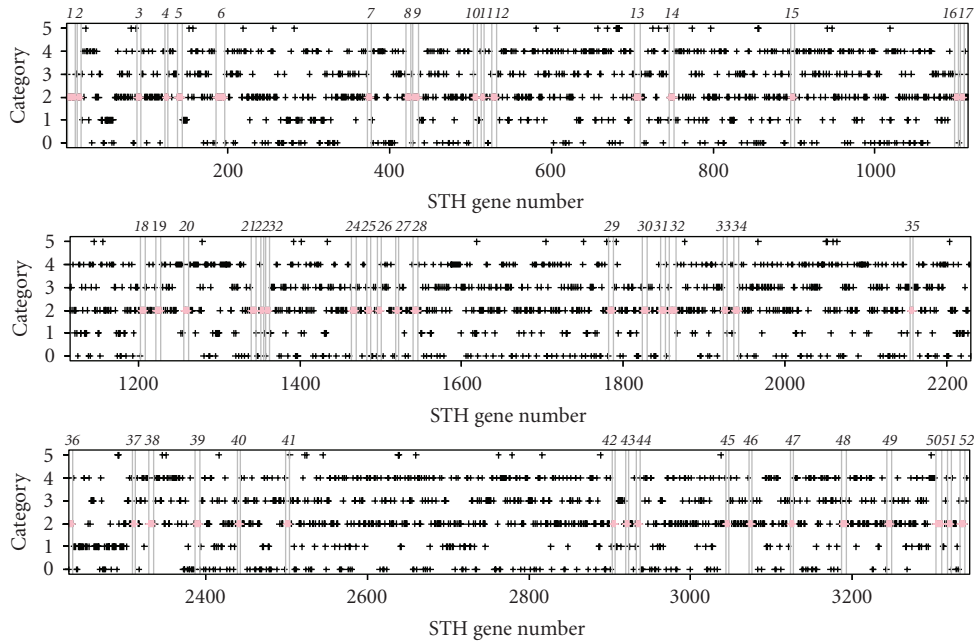

Figure 2: Plots of the location and category of the *Symbiobacterium* protein-coding genes. *X*-axis: STH gene number. *Y*-axis: 0: category "No hit" (*Symbiobacterium*-specific genes); 1: category "*Symbiobacterium*" (multiple copied genes); 2: category "Clostridia;" 3: category "Bacilli"; 4: categories "Actinobacteria," "Aquificae," "Bacteroidetes/Chlorobi," "Chloroflexi," "Cyanobacteria," "*Deinococcus-Thermus,*" "Dictyoglomi," "Fibrobacteres/Acidobacteria," "Proteobacteria," "*Thermobaculum,*" "Thermotogae," and "Other bacteria"; 5: category "Archaea/Eukaryota." The italicized numbers indicate 52 clusters (pink) containing 5 or more consecutive genes belonging to the category "Clostridia."

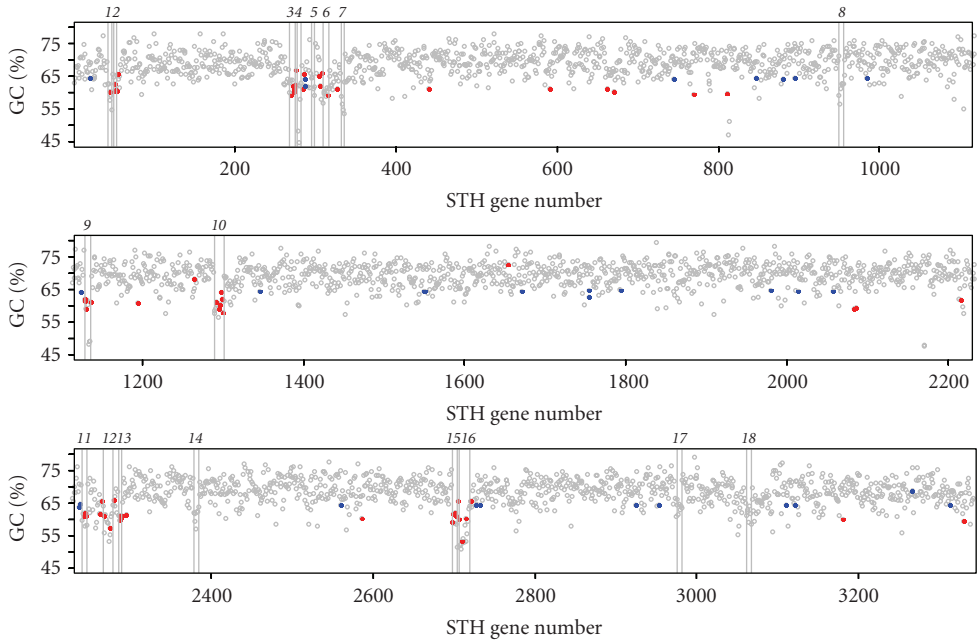

Figure 3: Plots of the location and GC content of the *Symbiobacterium* protein-coding genes. *X*-axis: STH gene number. *Y*-axis: GC content (%) of gene. Red indicates the putative transposase-coding genes, and blue indicates the group II intron-coding maturase genes. The italicized numbers indicate 18 low GC content regions containing 5 or more consecutive genes whose GC contents are below 65%.

do not overlap with the 52 vertically inherited clusters (Supplementary Table 1). On the basis of the KEGG gene cluster database, we found 12 gene clusters in the 18 low GC content regions (Supplementary Table 2).

Approximately 25% of the 3338 *Symbiobacterium* protein-coding genes belonged to categories consisting of organisms phylogenetically distant from *Symbiobacterium* (Figure 1), suggesting that *Symbiobacterium* frequently acquired genes during evolution. The proportion of horizontally transferred genes in the *Symbiobacterium* genome is strongly suggested to be the highest among bacteria [15]. These putative horizontally transferred genes are scattered across the genome of *S. thermophilum* (Figure 2). In addition, considering the species diversification of Bacilli and Clostridia, it is suggested that the categories "Bacilli" and "Clostridia" include not only vertically inherited genes but also horizontally transferred genes.

Transposase genes are generally used as markers of transposable genetic elements [16]. Most transposase-coding genes flank horizontally transferred genes [13]. *S. thermophilum* has 66 putative transposase-coding genes, of which 38 (58%) are located in the low GC content regions (P-value = 1.3×10^{-99}; Pearson's chi-square test) (Figure 3), suggesting that *Symbiobacterium* has a similar silencing system as that of *Salmonella* [17, 18]. In the silencing system, a histone-like nucleoid structuring (H-NS) protein binds to the region with a low GC content. Similar functional (H-NS) proteins were reported in *Mycobacterium* and *Pseudomonas* [19, 20]. If *Symbiobacterium* also has such proteins that bind the low GC content regions, the expression of the transposable elements located in these regions might be inhibited. As mentioned above, most transposable elements belong to the category "*Symbiobacterium,*" which is consistent with the fact that the genes of this category have lower GC contents than those of the other categories (Supplementary Figure 1). The regions consisting of low GC content genes cannot be explained by the directional mutation pressure or amelioration of bacterial genomes [21, 22]. Interestingly, although H-NS proteins bind the low GC content regions in *Mycobacterium*, *Pseudomonas*, and *Salmonella* [17–20], the H-NS protein of *Escherichia coli* does not specifically bind only these regions [23].

In addition, *S. thermophilum* has 30 group II intron-encoding maturase genes. Group II introns are transposable elements [24] that encode maturase as an intron-specific splicing factor [25]. The GC content of each maturase gene is approximately 65% (Figure 3). These maturase genes are classified in "*Symbiobacterium,*" on the basis of amino acid sequence similarity. In contrast to the transposase genes, the group II intron-encoding maturase genes are not located in the 18 low GC content regions (Figure 3). If *Symbiobacterium* has both an H-NS protein binding the low GC content regions and a gene silencing system similar to *Mycobacterium*, *Pseudomonas*, and *Salmonella*, these maturases could be activated and the group II introns could be transposed to the *Symbiobacterium* genome. Of course, it is also possible that this transposition of the group II introns is inhibited by another gene silencing system.

It is suggested that *Symbiobacterium* has gained many DNA fragments from phylogenetically distant organisms during the early stage of evolution in the Firmicutes (consisting of Bacilli and Clostridia). As the *Symbiobacterium* genes of all categories have a high GC content (Supplementary Figure 1), it can be concluded that, after acquiring genes, *Symbiobacterium* increased the GC content of the horizontally transferred genes and thereby maintained a genome with a high GC content.

In contrast to the *Symbiobacterium* genome, the *Fusobacterium* (phylogenetically closely related to Firmicutes) genome has a low GC content (27%) [13]. It is suggested that *Fusobacterium* has gained many genes from phylogenetically distant organisms [13]. In the course of evolution, *Fusobacterium* has probably decreased the GC content of the horizontally acquired genes and maintained a genome with a low GC content.

Does *Symbiobacterium* benefit from maintaining a genome with a high GC content? Considering that CO_2 is the major growth factor of *Symbiobacterium*, its symbiotic partners may not be limited to *Geobacillus*. *Symbiobacterium* is widespread in different natural environments [26, 27]. The difference in the genome base compositions between *Symbiobacterium* and its symbiotic partners may lead to a decrease in the frequency of a homologous recombination between the 2 genomes. For example, the 5 sequenced chromosomal genomes of *Geobacillus* have a GC content ranging from 42.8% to 52.5% (http://insilico.ehu.es/oligoweb/).

In addition, homologous recombination is generally effective for adaptive evolution [11]. However, if the population density is low or the recombining population is rare in the environment, adaptive evolution is hampered [11]. Considering the wide distribution of *Symbiobacterium* in natural environments, the population size of *Symbiobacterium* may be adequately large, suggesting that homologous recombination between the *Symbiobacterium* strains and different symbiotic partners may be effective for adaptive evolution. Thus, it is hypothesized that *Symbiobacterium* has maintained its extreme genome composition to avoid homologous recombination between its genome and the genomes of different species and to promote homologous recombination between its genome and the genomes of the same species (or genus).

Acknowledgment

The authors thank Professor Teruhiko Beppu for the helpful comments and critical review of the paper.

References

[1] K. Ueda and T. Beppu, "Lessons from studies of *Symbiobacterium thermophilum*, a unique syntrophic bacterium," *Bioscience, Biotechnology and Biochemistry*, vol. 71, no. 5, pp. 1115–1121, 2007.

[2] H. Nishida, T. Beppu, and K. Ueda, "*Symbiobacterium* lost carbonic anhydrase in the course of evolution," *Journal of Molecular Evolution*, vol. 68, no. 1, pp. 90–96, 2009.

[3] K. Ueda, A. Yamashita, J. Ishikawa et al., "Genome sequence of *Symbiobacterium thermophilum*, an uncultivable bacterium that depends on microbial commensalism," *Nucleic Acids Research*, vol. 32, no. 16, pp. 4937–4944, 2004.

[4] T. O. Watsuji, T. Kato, K. Ueda, and T. Beppu, "CO_2 supply induces the growth of *Symbiobacterium thermophilum*, a syntrophic bacterium," *Bioscience, Biotechnology and Biochemistry*, vol. 70, no. 3, pp. 753–756, 2006.

[5] G. Ding, Z. Yu, J. Zhao et al., "Tree of life based on genome context networks," *PLoS One*, vol. 3, no. 10, Article ID e3357, 2008.

[6] E. P. C. Rocha and E. J. Feil, "Mutational patterns cannot explain genome composition: are there any neutral sites in the genomes of bacteria?" *PLoS Genetics*, vol. 6, no. 9, Article ID e1001104, 2010.

[7] F. Hildebrand, A. Meyer, and A. Eyre-Walker, "Evidence of selection upon genomic GC-content in bacteria," *PLoS Genetics*, vol. 6, no. 9, Article ID e1001107, 2010.

[8] R. Hershberg and D. A. Petrov, "Evidence that mutation is universally biased towards AT in bacteria," *PLoS Genetics*, vol. 6, no. 9, Article ID e1001115, 2010.

[9] J. P. Gogarten and J. P. Townsend, "Horizontal gene transfer, genome innovation and evolution," *Nature Reviews Microbiology*, vol. 3, no. 9, pp. 679–687, 2005.

[10] E. V. Koonin, K. S. Makarova, and L. Aravind, "Horizontal gene transfer in prokaryotes: quantification and classification," *Annual Review of Microbiology*, vol. 55, pp. 709–742, 2001.

[11] B. R. Levin and O. E. Cornejo, "The population and evolutionary dynamics of homologous gene recombination in bacteria," *PLoS Genetics*, vol. 5, no. 8, Article ID e1000601, 2009.

[12] J. M. Smith, N. H. Smith, M. O'Rourke, and B. G. Spratt, "How clonal are bacteria?" *Proceedings of the National Academy of Sciences of the United States of America*, vol. 90, no. 10, pp. 4384–4388, 1993.

[13] A. Mira, R. Pushker, B. A. Legault, D. Moreira, and F. Rodríguez-Valera, "Evolutionary relationships of *Fusobacterium nucleatum* based on phylogenetic analysis and comparative genomics," *BMC Evolutionary Biology*, vol. 4, article no. 50, 2004.

[14] C. A. Fuchsman and G. Rocap, "Whole-genome reciprocal BLAST analysis reveals that *Planctomycetes* do not share an unusually large number of genes with *Eukarya* and *Archaea*," *Applied and Environmental Microbiology*, vol. 72, no. 10, pp. 6841–6844, 2006.

[15] Y. Nakamura, T. Itoh, H. Matsuda, and T. Gojobori, "Biased biological functions of horizontally-transferred genes in prokaryotic genomes," *Nature Genetics*, vol. 36, no. 7, pp. 760–766, 2004.

[16] M. G. I. Langille, W. W. L. Hsiao, and F. S. L. Brinkman, "Detecting genomic islands using bioinformatics approaches," *Nature Reviews Microbiology*, vol. 8, no. 5, pp. 373–382, 2010.

[17] S. Lucchini, G. Rowley, M. D. Goldberg, D. Hurd, M. Harrison, and J. C. Hinton, "H-NS mediates the silencing of laterally acquired genes in bacteria.," *PLoS Pathogens*, vol. 2, no. 8, article e81, 2006.

[18] W. W. Navarre, S. Porwollik, Y. Wang et al., "Selective silencing of foreign DNA with low GC content by the H-NS protein in *Salmonella*," *Science*, vol. 313, no. 5784, pp. 236–238, 2006.

[19] C. -S. Yun, C. Suzuki, K. Naito et al., "Pmr, a histone-like protein H1 (H-NS) family protein encoded by the IncP-7 plasmid pCAR1, is a key global regulator that alters host function," *Journal of Bacteriology*, vol. 192, no. 18, pp. 4720–4731, 2010.

[20] B. R. G. Gordon, Y. Li, L. Wang et al., "Lsr2 is a nucleoid-associated protein that targets AT-rich sequences and virulence genes in *Mycobacterium tuberculosis*," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 107, no. 11, pp. 5154–5159, 2010.

[21] J. G. Lawrence and H. Ochman, "Amelioration of bacterial genomes: rates of change and exchange," *Journal of Molecular Evolution*, vol. 44, no. 4, pp. 383–397, 1997.

[22] N. Sueoka, "On the genetic basis of variation and heterogeneity of DNA base composition," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 48, pp. 582–592, 1962.

[23] T. Oshima, S. Ishikawa, K. Kurokawa, H. Aiba, and N. Ogasawara, "*Escherichia coli* histone-like protein H-NS preferentially binds to horizontally acquired DNA in association with RNA polymerase," *DNA Research*, vol. 13, no. 4, pp. 141–153, 2006.

[24] F. Michel and J. L. Ferat, "Structure and activities of group II introns," *Annual Review of Biochemistry*, vol. 64, pp. 435–461, 1995.

[25] M. Matsuura, J. W. Noah, and A. M. Lambowitz, "Mechanism of maturase-promoted group II intron splicing," *EMBO Journal*, vol. 20, no. 24, pp. 7259–7270, 2002.

[26] T. Sugihara, T. O. Watsuji, S. Kubota et al., "Distribution of *Symbiobacterium thermophilum* and related bacteria in the marine environment," *Bioscience, Biotechnology and Biochemistry*, vol. 72, no. 1, pp. 204–211, 2008.

[27] K. Ueda, M. Ohno, K. Yamamoto et al., "Distribution and diversity of symbiotic thermophiles, *Symbiobacterium thermophilum* and related bacteria, in natural environments," *Applied and Environmental Microbiology*, vol. 67, no. 9, pp. 3779–3784, 2001.