Chapter 3

Lateral genes in the genomic DNA sequence

3.1 The correlation between the G+C content of the genomic DNA sequences and the G+C content at the third codon position (GC3 content) for all ORFs of the genomic DNA sequences

As shown in Fig. 1, the scores of the G+C content (GC content) of twenty-three genomic DNA sequences show the range to be between approximately 25 and 70%. There was a strong positive correlation, however, reported between the GC content of the genomic sequence and the GC content at each codon position in all ORFs of each genomic DNA sequence (*GC3-plot* named in ref 55) [56] . Nevertheless, the G+C content at the third codon position(GC3 content) in all ORFs demonstrated the best correlation as shown in Fig. 2, which was similar to the result described in ref 57. Consequently, the use of the GC3 content as a gene index of the gene is suggested to lead to better results, since the genes are thus better patterned and standardized.

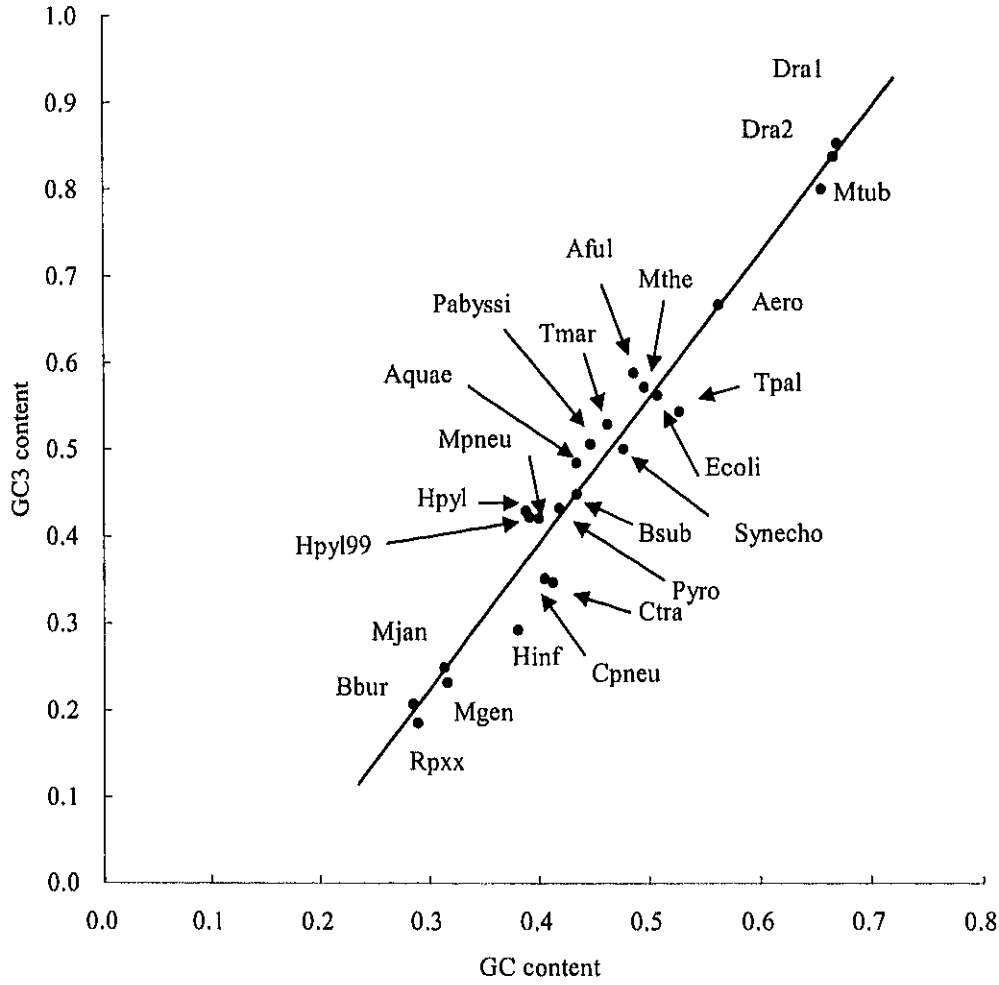


Figure 1 Plot of the G+C content between the complete genomic DNA sequences (GC content) and the third codon position of all the ORFs (GC3 content).
Abbreviations are as listed in Table 1. A regression line is fitted to the data.

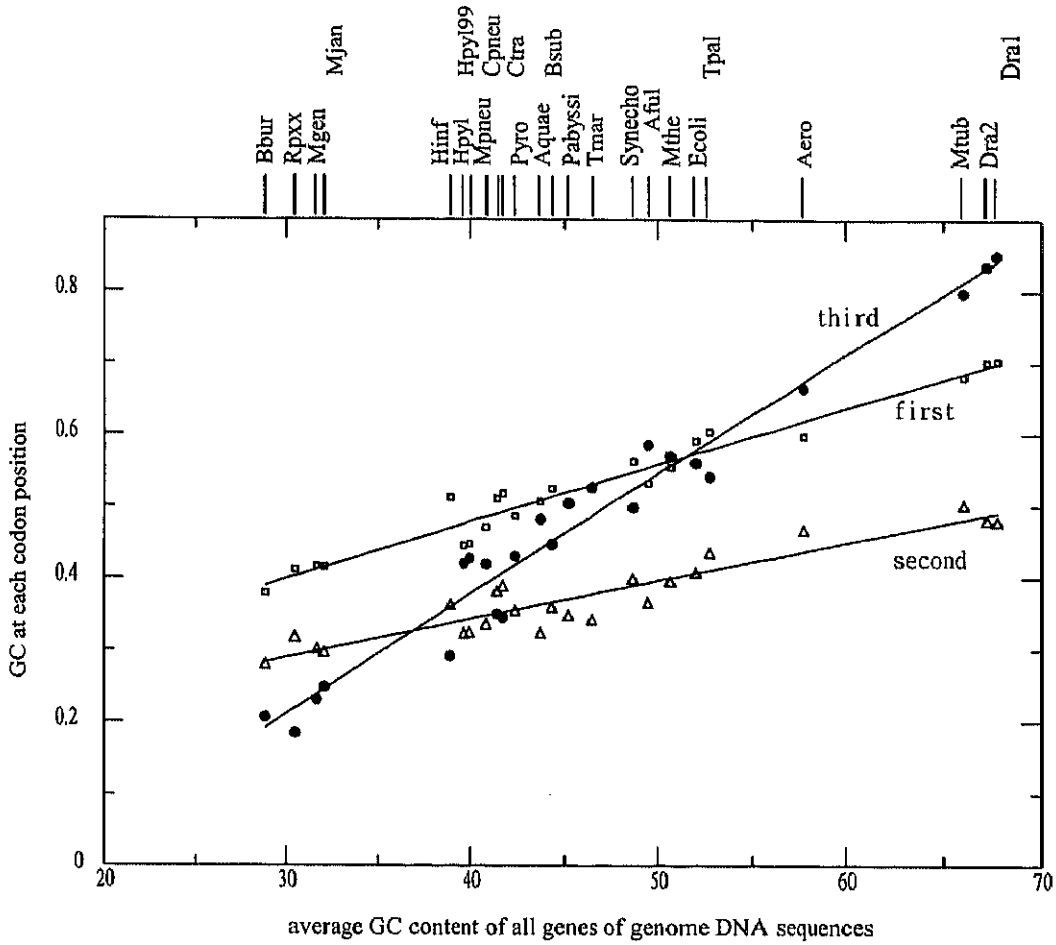


Figure 2 Plots of the average GC contents at each codon position for all genes against the G+C contents of the genomic DNA sequences used in this analysis.

The regression lines were fitted to the data. Open squares (□), open triangles (△) and filled circles (●) indicate the average G+C content of the first, second and third codon positions, respectively. Abbreviations are as listed in Table 1.

3.2 The correlation between the G+C content of the genomic DNA sequences and the G+C content at the third position of the three types of codons found in orthologous genes

Codons compared between orthologous genes were classified into three categories, IA, DC, and IC. IC indicates the identical codons coding the same amino acids; DC indicates the different codons coding the different amino acids; and IA indicates the synonymous codons coding the same amino acids (synonymous codons). As shown in Fig. 3, D0 indicates the differences of GC content between species 1 and species 2 and D1, D2 indicate the bias of the GC3 content between species 1 and the value calculated using the regression line at the GC content of species 1 and that of species 2 and the value calculated using the regression line at the GC content of species 2. And D12 showed the difference of GC3 content between species 1 and species 2.

Fig. 4 shows the correlation between the GC3 content of IC, DC, and IA in the orthologous genes among *Ph* and *Pa* and the GC content of the genomic DNA sequences. The CG3 content of IC is not taken into consideration since the GC3 content does not vary between the two species. The deviation of the GC3 content of DC was similar to the expectation value of the GC content of the genomic DNA sequences. The deviation of the GC3 content of IA was the largest. Therefore, the GC3 content of the synonymous codon in the orthologous genes was suggested to distinguish lateral gene from the stable genes of the genomic DNA sequence.

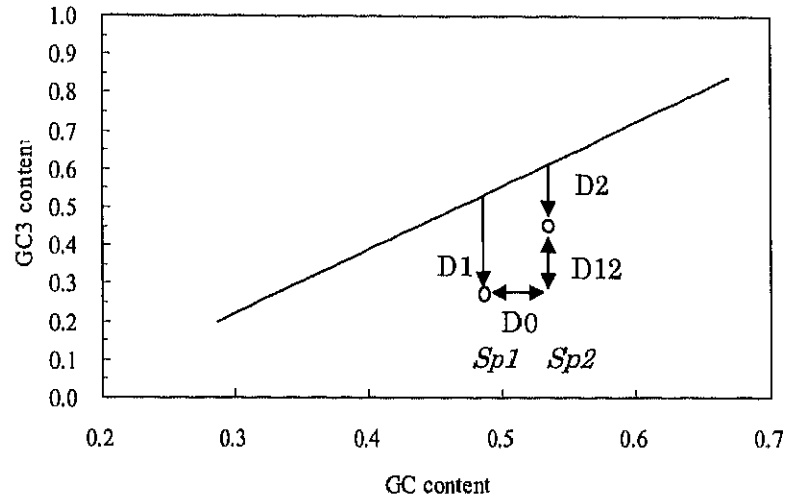
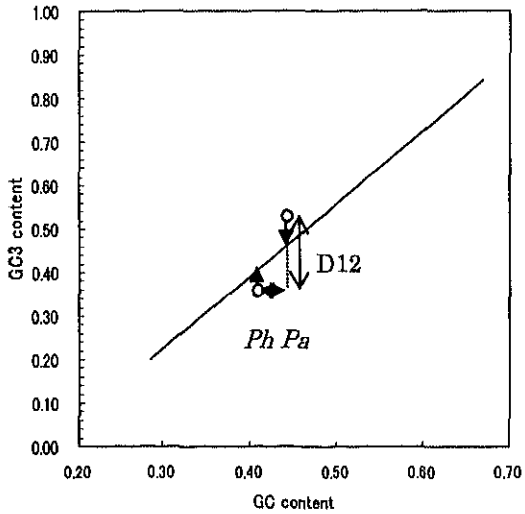


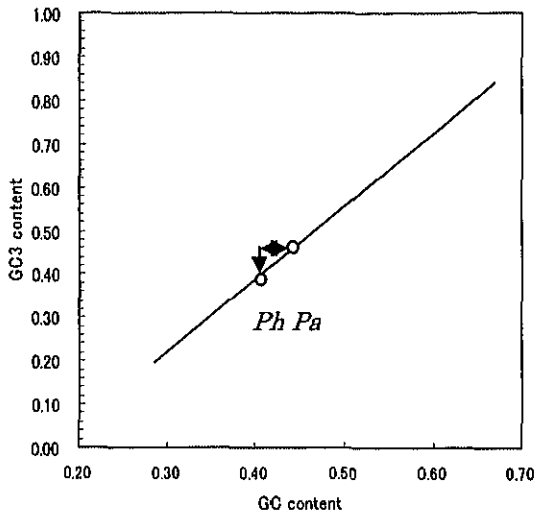
Figure 3 Illustration of the quantities of the differences of GC3 content

The circles represent the GC3 content for species 1 (*Sp1*) and species 2 (*Sp2*), as labeled. D0 represents the difference of the GC content between *Sp1* and *Sp2*. D1 and D2 indicate the difference of GC3 content in the overall GC3 content between *Sp1* and *Sp2* from the *GC3-plot*, respectively. D12 indicates the difference of GC3 content between *Sp1* and *Sp2*.

(a) IA



(b) DC



(c) IC

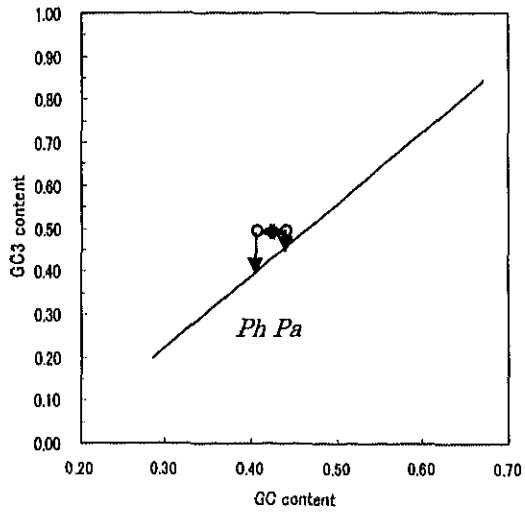


Figure 4 Separate plots of the GC3 content of IA, DC and IC, for *Ph/Pa* comparison in *GC3-plot*.

Ph is species 1 and *Pa* is species 2. (a) IA category. (b) DC category - only D0 and D12 are shown. (c) IC category -D12 is not shown since D12 is zero by definition.

3.3 Distribution of the G+C content at the third codon position of synonymous codons in the orthologous genes of *Pyrococcus horikoshii* OT3

Fig. 5 plots the distribution of the GC3 content of synonymous codons in the orthologous genes against the position of the genomic DNA sequence of *Ph.* The average of the GC3 content of synonymous codon in the orthologous genes was 0.365. The genes having synonymous codons with the higher GC3 content were located particularly in the region from approximately 300 Kbp to 420 Kbp (high GC3 content region). And the genes having synonymous codons with the lower GC3 content were located in the region from approximately 1,320 Kbp to 1,400 Kbp (low GC3 content region).

As shown in Fig. 6 (a), the distribution of GC3 content of orthologous genes in *Ph* was regarded as a Gaussian distribution having an average and standard deviation of 0.365 and 0.0709, respectively. It was reported that 780 genes in the DNA sequence of *Escherichia coli*, which was corresponded to approximately 30% of the whole genomic sequence, were analyzed by Fractional Correspondence Analysis and the Dynamic Clustering Method and the genes were classified into three classes [53] . The 111 genes constituting the horizontally inherited transfer genes comprised the class with the smallest codon bias among the three classes [53] . Therefore, the score of 0.142 with respect to the total genes analyzed suggests these genes to be transferred in this DNA sequence. When this value is applied to the distribution of GC3 content of the orthologous genes, the coefficient of the standard deviation is approximately 1.5. Thus, the values of the GC3 content were 0.471 and 0.233. For the distribution of the GC3 content of putative lateral genes, it might be likely that the curve would increase and decrease near both edges of the distribution. In addition, patterns of the GC3 content of the synonymous codons in both regions, as shown in Fig. 6 (b), were different from that of all orthologous genes in the genomic DNA sequence. Therefore, by exploring such values that the distribution curve would increase and decrease near both edges of the distribution curve for GC3 content of orthologous genes, we chose the two values, both 0.475 and 0.245 to distinguish the genes of high GC3 content and low GC3 content, respectively, from the orthologous genes. These values were then applied as thresholds for distinguishing lateral genes from the stable genes of *Ph.* In the genomic DNA sequence of *Ph*, there are 78 orthologous genes, whose GC3 content of the synonymous codons is more

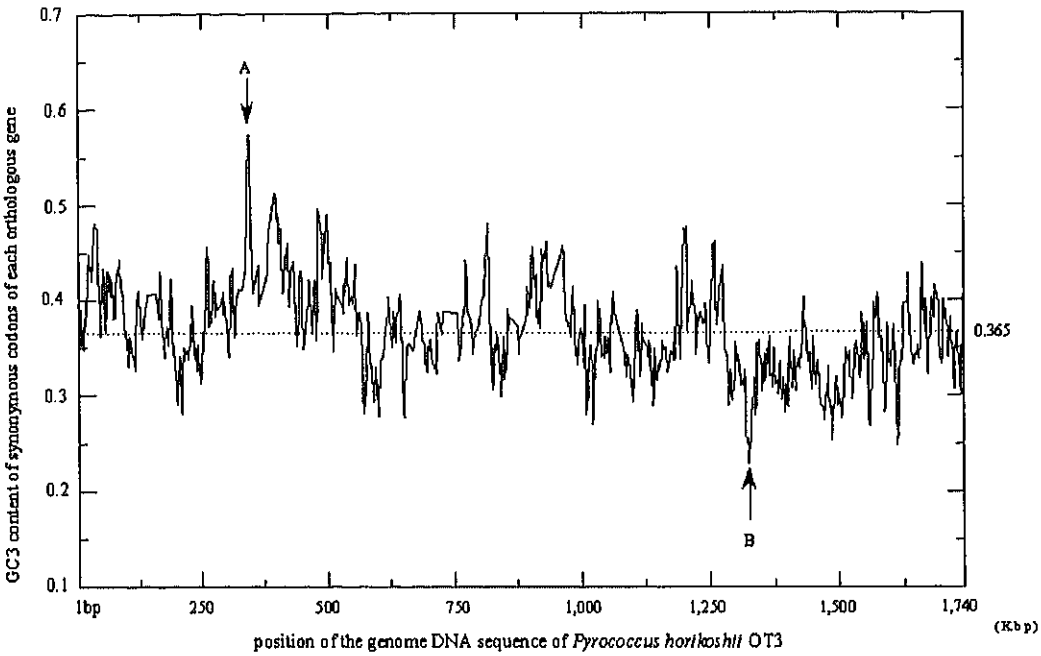
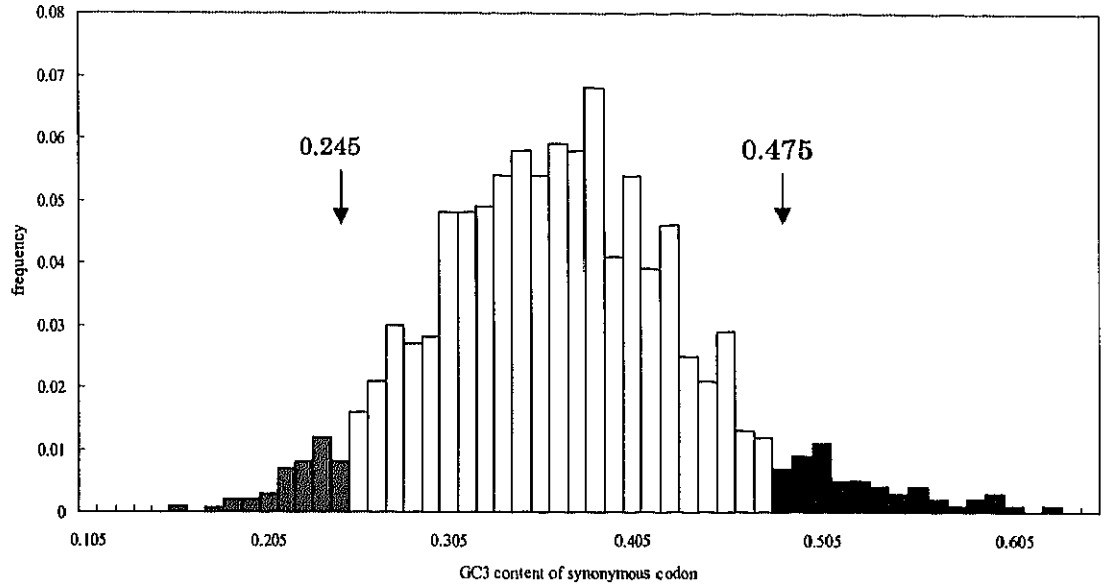


Figure 5 Plot of the GC3 content of synonymous codons of orthologous genes against the genomic DNA sequence of *Ph*.

A and B indicate the high GC3 and the low GC3 content region, respectively. The overall average of the GC3 content of synonymous codons is 0.365.

(a) Histogram of GC3 content of synonymous codon of all orthologous genes



(b) comparison of distribution of GC3 content of synonymous codon

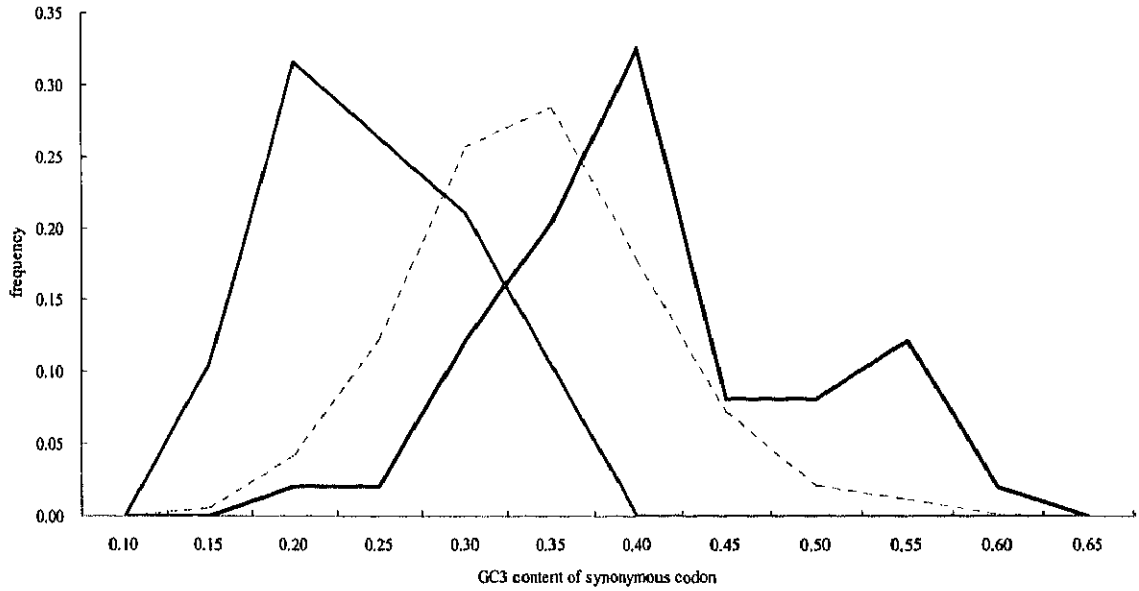


Figure 6 Distribution of the GC3 content of synonymous codons of all orthologous genes in *Ph.*

(a) The histogram of GC3 content of synonymous codons of all orthologous genes.
(b) The distribution of GC3 content of synonymous codons of all orthologous genes, the orthologous genes in the region from 300 to 402 Kbp, and those in the region from 1,320 to 1,340 Kbp, as represented by a gray dashed line, a black bold line and a gray bold line, respectively.

than or equal to 0.475. Sixteen genes were located in the high GC3 content region in Fig. 5, region A. In 48 of the orthologous genes, however, the GC3 content of synonymous codons is less than or equal to 0.245. Thirteen genes were located in the low GC3 content region in Fig. 5, region B.

3.4 Characteristics of the high and low GC3 content regions

By analysis of 126 orthologous genes, 40 orthologous genes were homologous to genes from Bacteria and Archaea, 24 genes from Eukarya and Archaea and 54 from Archaea. Eight orthologous genes had homologous genes from all three domains, Archaea, Bacteria and Eukarya as shown in Fig. 7.

In the high GC3 content region shown in Fig. 5, region A, 63 genes were orthologous genes, 49% and 8% of which showed homology to the genes from Bacteria and Archaea and to those of Eukarya and Archaea, respectively. Twelve of 16 orthologous genes, whose GC3 content of synonymous codons is greater than or equal to 0.475, were homologous to the genes from Bacteria and Archaea. In the low GC3 content region shown in Fig. 5, region B, 69 genes were orthologous genes, 41% and 25% of which were homologues to the genes from Eukarya and Archaea and from Bacteria and Archaea, respectively. Twelve of 13 orthologous genes, whose GC3 content of synonymous codons is less than or equal to 0.245, were homologous to the genes from Archaea and Eukarya.

Four orthologous genes in the high GC3 content region, region A as shown in Table 3 (a), were classified into the category of cell structure, because they were homologous to genes related to cell wall or lipopolysaccharides. They also showed high homology to genes related to the cell envelope or biosynthesis of surface polysaccharides and lipopolysaccharides from *Streptococcus*. On the other hand, the organization of the operon related to this function is known to differ among *Streptococcus thermophilus* species [58] . The 13.6 Kbp *epsL-IS981SC* region of *Streptococcus thermophilus* CNRZ386 is closely related to the functionally similar sequence of *Lactococcus lactis* NIZB40. The IS elements (ISS*1* and IS*981*) shared more than 98% identity to the homologous ISS*1* and IS*981* from *L. lactis*. In addition, probes isolated from this region hybridized with various DNA fragments of *L. lactis* and the GC content in this region showed variability. This suggests that the entire *epsL-IS981SC* region was transferred from *L. lactis* [58] . The four orthologous genes had a bias of GC3 content of synonymous codons and were homolo-

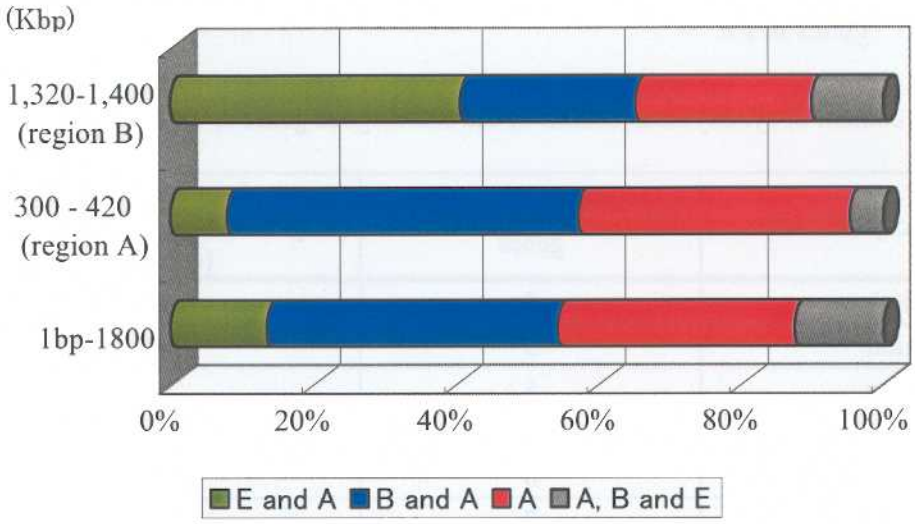


Figure 7 Distribution of the three domains of genes homologous to the orthologous genes in *Pyrococcus horikoshii* OT3.
Characters E, A and B indicate Eukarya, Archaea and Bacteria, respectively.

Table 3 Number of the GC3 content of synonymous codons of orthologous genes in *Ph*

(a) GC3 content of synonymous codons of the orthologous genes is 0.475 or greater.

category of function	number of the orthologous genes	number of the orthologous genes in the 300-420 Kbp (region A)
transcription	3	0
translation	5	0
replication	5	1
cell process	5	1
cell structure	8	4
metabolism	7	2
others	0	0
hypothetical protein	45	8

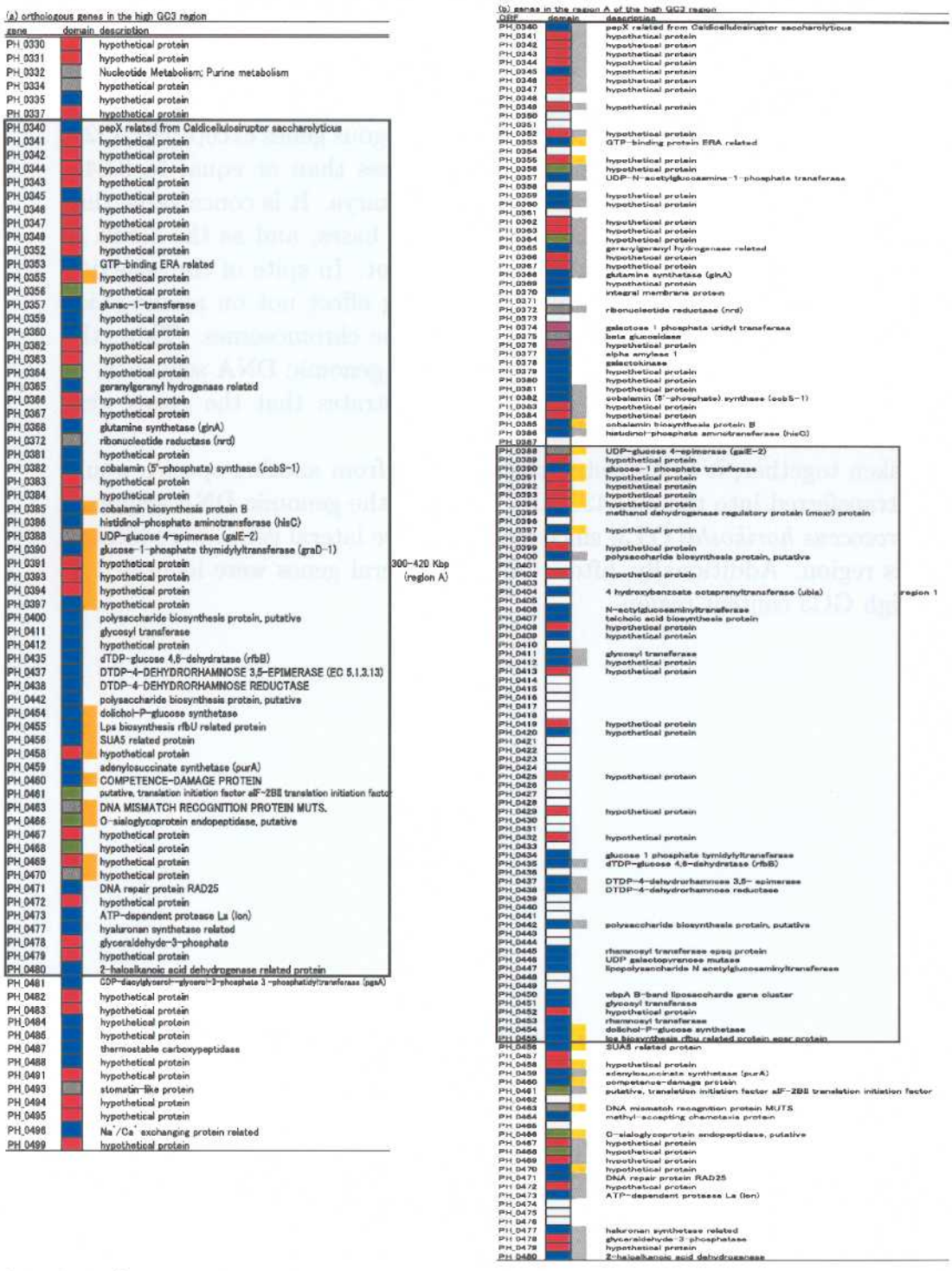
(b) GC3 content of synonymous codons of the orthologous genes is 0.245 or less.

category of function	number of the orthologous genes	number of the orthologous genes in the 1,320-1,400 Kbp (region B)
transcription	2	0
translation	11	6
replication	0	0
cell process	3	1
cell structure	1	0
metabolism	5	1
others	1	0
hypothetical protein	25	5

gous to the genes from Bacteria and Archaea in Fig. 8 (b), region 1. I suggest that they are likely to be lateral genes.

In the low GC3 content region in Fig. 8 (c), region B, six orthologous genes were classified into the category of translation as shown in Table 2 (b). Among these six, five were homologous to genes coding for ribosomal proteins. As shown in Fig. 8 (c), all of the orthologous genes except PH_1529, whose GC3 content of synonymous codons is less than or equal to 0.245, were homologous to genes from Archaea and Eukarya. It is conceivable that mutations had been occurring in the nucleotide bases, and as the result of mutation, amino acids were either changed or not. In spite of the selection pressure in evolution, changes have been taking effect not on amino acids or proteins but rather on nucleotide bases or the chromosomes. Thus, the smallest deviation from the GC3 content of the genomic DNA sequence, in comparison with that of other regions, demonstrates that the genes were mostly unaffected in evolution.

Taken together, it is suggested that a region from another species would have transferred into the 300-420 Kbp region of the genomic DNA sequence of *Pyrococcus horikoshii* OT3, since eight putative lateral genes were located in this region. Additionally, fifteen putative lateral genes were identified in the high GC3 content region.


(a) orthologous genes in the high GC3 region
gene domain description
PH_0330 hypothetical protein
PH_0331 hypothetical protein
PH_0332 Nucleotide Metabolism; Purine metabolism
PH_0334 hypothetical protein
PH_0335 hypothetical protein
PH_0337 hypothetical protein
PH_0340 pepX related from Caldicellulosiruptor saccharolyticus
PH_0341 hypothetical protein
PH_0342 hypothetical protein
PH_0344 hypothetical protein
PH_0343 hypothetical protein
PH_0345 hypothetical protein
PH_0346 hypothetical protein
PH_0347 hypothetical protein
PH_0348 hypothetical protein
PH_0352 hypothetical protein
PH_0353 GTP-binding ERA related
PH_0355 hypothetical protein
PH_0356 hypothetical protein
PH_0357 glunac-1-transferase
PH_0359 hypothetical protein
PH_0360 hypothetical protein
PH_0362 hypothetical protein
PH_0363 hypothetical protein
PH_0364 hypothetical protein
PH_0365 geranylgeranyl hydrogenase related
PH_0366 hypothetical protein
PH_0367 hypothetical protein
PH_0368 glutamine synthetase (glnA)
PH_0372 ribonucleotide reductase (nrd)
PH_0381 hypothetical protein
PH_0382 cobalamin (5'-phosphate) synthase (cobS-1)
PH_0383 hypothetical protein
PH_0384 hypothetical protein
PH_0385 cobalamin biosynthesis protein B
PH_0386 histidinol-phosphate aminotransferase (hisC)
PH_0388 UDP-glucose 4-epimerase (galE-2)
PH_0390 glucose-1-phosphate thymidylyltransferase (graD-1)
PH_0391 hypothetical protein
PH_0393 hypothetical protein
PH_0394 hypothetical protein
PH_0397 hypothetical protein
PH_0400 polysaccharide biosynthesis protein, putative
PH_0411 glycosyl transferase
PH_0412 hypothetical protein
PH_0435 dTDP-glucose 4,6-dehydratase (rfbB)
PH_0437 DTDP-4-DEHYDRORHAMNOSE 3,5-EPIMERASE (EC 5.1.3.13)
PH_0438 DTDP-4-DEHYDRORHAMNOSE REDUCTASE
PH_0442 polysaccharide biosynthesis protein, putative
PH_0454 dolichol-P-glucose synthetase
PH_0455 Lps biosynthesis rfbU related protein
PH_0456 SUA5 related protein
PH_0458 hypothetical protein
PH_0459 adenylosuccinate synthetase (purA)
PH_0460 COMPETENCE-DAMAGE PROTEIN
PH_0461 putative, translation initiation factor aIF-2BII translation initiation factor
PH_0463 DNA MISMATCH RECOGNITION PROTEIN MUTS.
PH_0466 O-sialoglycoprotein endopeptidase, putative
PH_0467 hypothetical protein
PH_0468 hypothetical protein
PH_0469 hypothetical protein
PH_0470 hypothetical protein
PH_0471 DNA repair protein RAD25
PH_0472 hypothetical protein
PH_0473 ATP-dependent protease La (lon)
PH_0477 hyaluronan synthetase related
PH_0478 glyceraldehyde-3-phosphate
PH_0479 hypothetical protein
PH_0480 2-haloalkanoic acid dehydrogenase related protein
PH_0481 CDP-diacylglycerol--glycerol-3-phosphate 3 -phosphatidyltransferase (pgsA)
PH_0482 hypothetical protein
PH_0483 hypothetical protein
PH_0484 hypothetical protein
PH_0485 hypothetical protein
PH_0487 thermostable carboxypeptidase
PH_0488 hypothetical protein
PH_0491 hypothetical protein
PH_0493 stomatin-like protein
PH_0494 hypothetical protein
PH_0495 hypothetical protein
PH_0496 Na+/Ca+ exchanging protein related
PH_0499 hypothetical protein
300-420 Kbp
(region A)
(b) genes in the region A of the high GC3 region
ORF domain description
PH_0340 pepX related from Caldicellulosiruptor saccharolyticus
PH_0341 hypothetical protein
PH_0342 hypothetical protein
PH_0343 hypothetical protein
PH_0344 hypothetical protein
PH_0345 hypothetical protein
PH_0346 hypothetical protein
PH_0347 hypothetical protein
PH_0348
PH_0349 hypothetical protein
PH_0350
PH_0351
PH_0352 hypothetical protein
PH_0353 GTP-binding protein ERA related
PH_0354
PH_0355 hypothetical protein
PH_0356 hypothetical protein
PH_0357 UDP-N-acetylglucosamine-1-phosphate transferase
PH_0358
PH_0359 hypothetical protein
PH_0360 hypothetical protein
PH_0361
PH_0362 hypothetical protein
PH_0363 hypothetical protein
PH_0364 hypothetical protein
PH_0365 geranylgeranyl hydrogenase related
PH_0366 hypothetical protein
PH_0367 hypothetical protein
PH_0368 glutamine synthetase (glnA)
PH_0369 hypothetical protein
PH_0370 integral membrane protein
PH_0371
PH_0372 ribonucleotide reductase (nrd)
PH_0373
PH_0374 galactose 1 phosphate uridyl transferase
PH_0375 beta glucosidase
PH_0376 hypothetical protein
PH_0377 alpha amylase 1
PH_0378 galactokinase
PH_0379 hypothetical protein
PH_0380 hypothetical protein
PH_0381 hypothetical protein
PH_0382 cobalamin (5'-phosphate) synthase (cobS-1)
PH_0383 hypothetical protein
PH_0384 hypothetical protein
PH_0385 cobalamin biosynthesis protein B
PH_0386 histidinol-phosphate aminotransferase (hisC)
PH_0387
PH_0388 UDP-glucose 4-epimerase (galE-2)
PH_0389 hypothetical protein
PH_0390 glucose-1-phosphate transferase
PH_0391 hypothetical protein
PH_0392 hypothetical protein
PH_0393 hypothetical protein
PH_0394 hypothetical protein
PH_0395 methanol dehydrogenase regulatory protein (moxr) protein
PH_0396
PH_0397 hypothetical protein
PH_0398
PH_0399 hypothetical protein
PH_0400 polysaccharide biosynthesis protein, putative
PH_0401
PH_0402 hypothetical protein
PH_0403
PH_0404 4 hydroxybenzoate octaprenyltransferase (ubia)
PH_0405
PH_0406 N-acetylglucosaminyltransferase
PH_0407 teichoic acid biosynthesis protein
PH_0408 hypothetical protein
PH_0409 hypothetical protein
PH_0410
PH_0411 glycosyl transferase
PH_0412 hypothetical protein
PH_0413 hypothetical protein
PH_0414
PH_0415
PH_0416
PH_0417
PH_0418
PH_0419 hypothetical protein
PH_0420 hypothetical protein
PH_0421
PH_0422
PH_0423
PH_0424
PH_0425 hypothetical protein
PH_0426
PH_0427
PH_0428
PH_0429 hypothetical protein
PH_0430
PH_0431
PH_0432 hypothetical protein
PH_0433
PH_0434 glucose 1 phosphate tymidylyltransferase
PH_0435 dTDP-glucose 4,6-dehydratase (rfbB)
PH_0436
PH_0437 DTDP-4-dehydrorhamnose 3,5- epimerase
PH_0438 DTDP-4-dehydrorhamnose reductase
PH_0439
PH_0440
PH_0441
PH_0442 polysaccharide biosynthesis protein, putative
PH_0443
PH_0444
PH_0445 rhamnosyl transferase epsq protein
PH_0446 UDP galactopyranose mutase
PH_0447 lipopolysaccharide N acetylglucosaminyltransferase
PH_0448
PH_0449
PH_0450 wbpA B-band liposaccharde gene cluster
PH_0451 glycosyl transferase
PH_0452 hypothetical protein
PH_0453 rhamnosyl transferase
PH_0454 dolichol-P-glucose synthetase
PH_0455 lps biosynthesis rfbu related protein epsc protein
PH_0456 SUA5 related protein
PH_0457
PH_0458 hypothetical protein
PH_0459 adenylosuccinate synthetase (purA)
PH_0460 competence-damage protein
PH_0461 putative, translation initiation factor aIF-2BII translation initiation factor
PH_0463 DNA mismatch recognition protein MUTS
PH_0464 methyl-accepting chemotaxis protein
PH_0465
PH_0466 O-sialoglycoprotein endopeptidase, putative
PH_0467 hypothetical protein
PH_0468 hypothetical protein
PH_0469 hypothetical protein
PH_0470 hypothetical protein
PH_0471 DNA repair protein RAD25
PH_0472 hypothetical protein
PH_0473 ATP-dependent protease La (lon)
PH_0474
PH_0475
PH_0476
PH_0477 hakronan synthetase related
PH_0478 glyceraldehyde-3-phosphatase
PH_0479 hypothetical protein
PH_0480 2-haloalkanoic acid dehydrogenase
region 1

(continued)

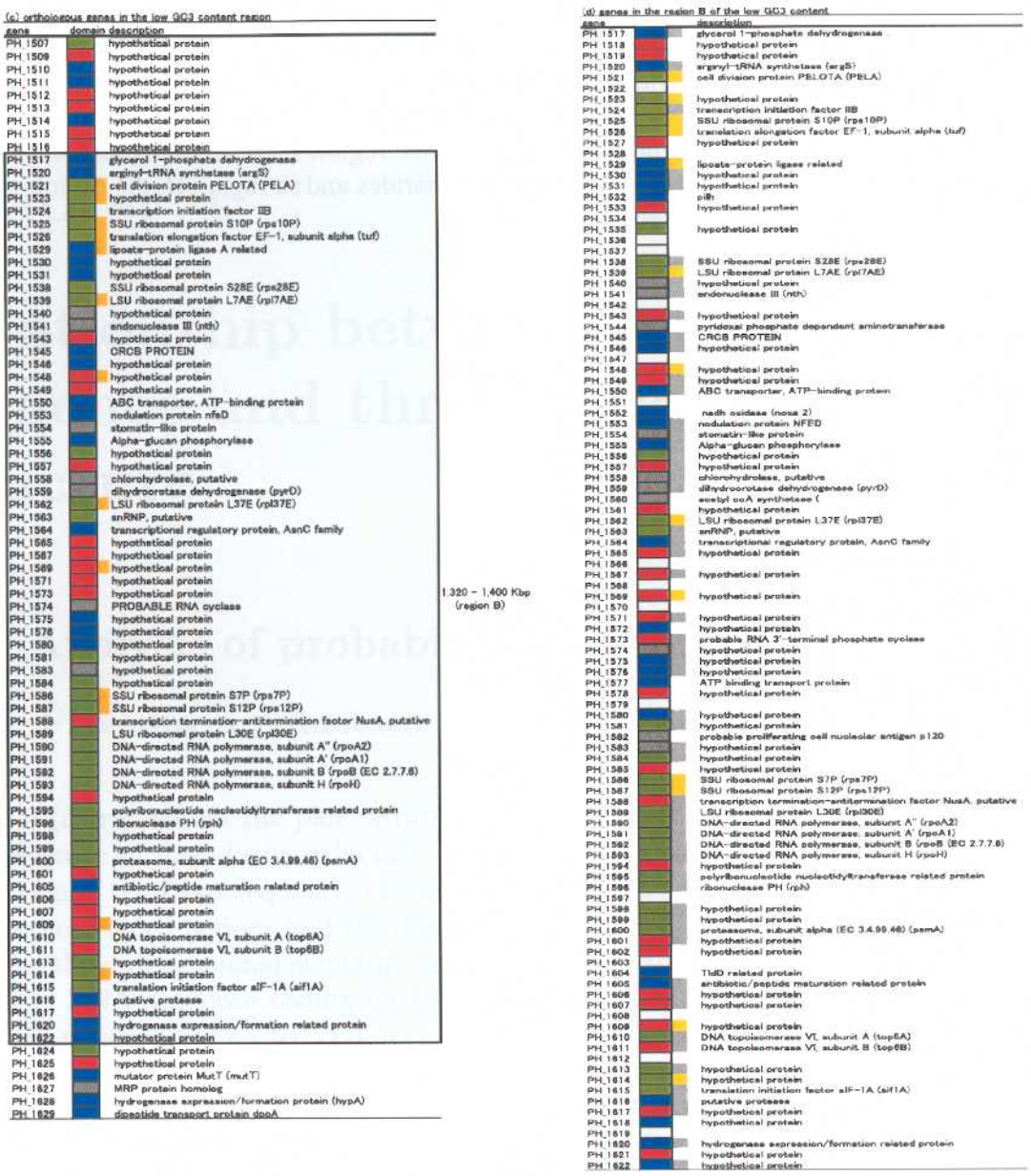


Figure 8 Gene maps of the high and low GC3 content region.

(a) and (b) show the high content region and (c) and (d) show the low content region. Red blocks indicate the homologues from Archaea. Blue ones indicate those from Bacteria and Archaea. Green ones indicate those from Eukarya and Archaea. Gray ones indicate those from Eukarya, Bacteria and Archaea. Violet ones indicate that from Eukarya and Bacteria. Yellow squares represent the genes having the higher GC3 content of

synonymous codons, greater than or equal to 0.475, in region A of the high GC3 content region and those having the lower GC3 content, lower than or equal to 0.245, in region B of the low GC3 content region. Gray squares represent the orthologous genes. In the region 1, the orthologous genes were related to cell-wall or lipopolysaccharides and in region B, the orthologous genes were related to translation.