Original Article

Journal of the Tsuruma HEALTH SCI. MED. KANAZAWA UNIV. Vol. 26
23～29 2002

The analysis of orthologous genes between closely related two pyrococcus species of archaebacteria

Horita Hiroshi Nakashima Hiroshi *

ABSTRACT

Orthologous genes from two archaebacteria, *Pyrococcus horikoshii* and *Pyrococcus abyssi*, were analyzed in terms of trinucleotide composition, nucleotide substitution and the G+C content at the third codon position. Genes were expressed as vectors and plotted as points in a composition space. The distribution of genes along the axis of the first principal component correlated with both the G+C content and the degree of sequence identity between pairs of orthologs. This result suggests the directional divergence from a common ancestral species yielded different G+C contents. Nucleotide substitutions from A/T of *P. horikoshii* to G/C of *P. abyssi* were most frequently observed. The substitutions of synonymous codons at the third position, accounting for 58% of all substitutions, were dominant. Investigation of the G+C content at the third codon position suggests that *P. horikoshii* undergoes a significant change resulting in a reduced G+C content, whereas that of *P. abyssi* remains unchanged.

KEY WORDS

Orthologous genes, Trinucleotide composition, Nucleotide substitution, G+C content

Introduction

Species-specific nucleotide compositions have been identified as a result of analyses of the ratio between observed and expected dinucleotide frequencies[1-3]. This is more clearly demonstrated when genes are analyzed in a composition space[4], where a sequence is converted to a composition vector and plotted as a point. The color display of the distribution of genes clearly indicated that genes are clustered around its average point in the composition space. To find some initial answers regarding the mechanism of species-specific nucleotide compositions, pairwise sequence alignments of orthologs from two mycoplasmas, *Mycoplasma genitalium* and *Mycoplasma pneumoniae*, were analyzed in a previous study[5]. The analysis indicated a directional gene distribution in a composition space suggesting a certain nucleotide substitution between two species. To elaborate on the previous study, the orthologous genes from two closely related pyrococcus species were analyzed in the study presented here, which focuses on the directional gene distribution in a composition space.

Materials and Methods

Orthologous genes

The genomic sequence data of two pyrococcus species have been determined and made available. This enabled us to obtain both the nucleotide and amino acid sequences of two species, *P. horikoshii* and *P. abyssi*, from the web site of the National Center for Biotechnology Information (http://www.ncbi.nlm.nih.gov/). These two pyrococcus species are extreme thermophilic archaea, with respective optimal growth temperatures of 98 and 103 degrees Celsius. A pair of orthologous genes from the two species were identified as follows. A homology search was carried out for a given protein sequence 'X' of *P. horikoshii* against all protein sequences of *P. abyssi*.

Central Clinical Laboratory, Kanazawa University Hospital
* School of Health Sciences, Faculty of Medicine, Kanazawa University

On the assumption that a given protein sequence 'Y' of *P. abyssi* would show the best match. Next, another homology search was performed for a given protein sequence 'Y' of *P. abyssi* against all protein sequences of P. horikoshii. If a given protein 'X' showed a best match with a 'Y' protein, 'X' and 'Y' were classified as a pair of orthologs. Homologous regions without any insertion or deletion in the alignment of orthologs were employed for sequence comparison. Nucleotide sequences corresponding to the protein sequences were also employed in the analysis. Sequences of less than 300 nucleotides were discarded to avoid relatively larger statistical errors in estimating trinucleotide composition. A total of 689 pairs of orthologous sequences, which included 667,506 pairs of nucleotides or 222,502 pairs of codons, were used for this study.

Analytical method

Trinucleotide compositions indicated a similar level of discrimination of genes between species as that of codon usage[6]. Therefore, orthologous genes were analyzed by using the trinucleotide composition of 32 independent components. Sixty-four kinds of trinucleotides were reduced to 32 by combining every complementary pair of trinucleotides (e.g. ApApA and TpTpT). The number of occurrences of trinucleotides is counted along a DNA sequence of a gene by shifting one base at a time, and this number is then converted to a composition (in percent) by dividing it by the total number of occurrences.

Subsequently, the nucleotide sequence of a given gene is converted to a trinucleotide composition vector of 32 components, which can be plotted as a point in a 32-dimensional composition space. For a clear display of the distribution of orthologous genes, they are plotted along the first principal component axis. A principal component analysis was then conducted separately for each set of genes of the two species. The coordinate for a gene along the first principal component axis was calculated by using the scalar product of the unit vector of the first principal component axis and the vector of the gene. The distribution of all orthologous genes of both species in this composition space was visualized by projecting them onto a two-dimensional plane, which was defined by two axes of the first principal components. The origin of the x-y coordinate system in the composition space was set at the average trinucleotide composition of all orthologous genes employed of both species.

The assignment of sequences to an organism was done as follows. The distance from a nucleotide sequence to two representative compositions was calculated in the normalized trinucleotide composition space, after which the nucleotide sequence was assigned to the organism with the shorter distance. The average composition of trinucleotides was expressed as a representative composition for a given organism. The distance between two nucleotide sequences i and j was calculated as follows.

$$d_{i,j} = (\Sigma (V_{i,k} - V_{j,k})^2)^{1/2} \quad (1)$$

Nucleotide sequence i was expressed as a trinucleotide composition vector in a normalized form,

$$V_{i,k} = (C_{i,k} - AV_k) / SD_k \quad (2)$$

Where, $V_{i,k}$ and $C_{i,k}$ are the normalized and real compositions of trinucleotides of the k-th component, and AV_k and SD_k are the average composition and the standard deviation for the whole dataset, respectively.

The shortest distance between two substitutions along the sequence was calculated. For example, comparing two sequences, ATGCAGC and ACGTGGC, yielded three substitutions at positions 2, 4 and 5, so the intervals between the substitutions became two and one, respectively. Then, one was added to the number of substitutions for intervals of both one and two. In this way, the occurrence of substitutions for the intervals from one to twenty residues was counted for all the pairwise alignments of orthologs.

The codons were classified into three groups according to the definition of Bellgard and Gojobori[7]. Group 1 codons, referred to as IA, are different but code for the identical amino acid. Group 2 codons, DA, are different and code for a different amino acid. Group 3 codons, IC, are identical codons. The average G+C content at the third position was calculated separately for the three groups of codons.

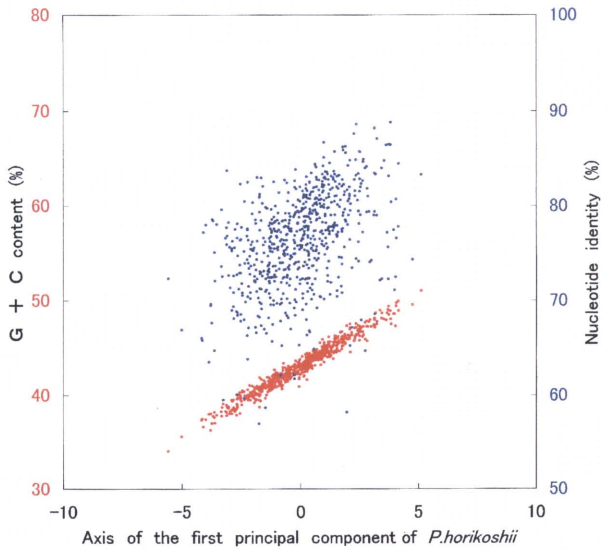

Figure 1. Plotting of genes of *P. horikoshii* projected on the first principal component axis against the G+C content of the genes (red dots on the left-hand scale), as well as against the sequence identity between the orthologs (blue dots on the right-hand scale).

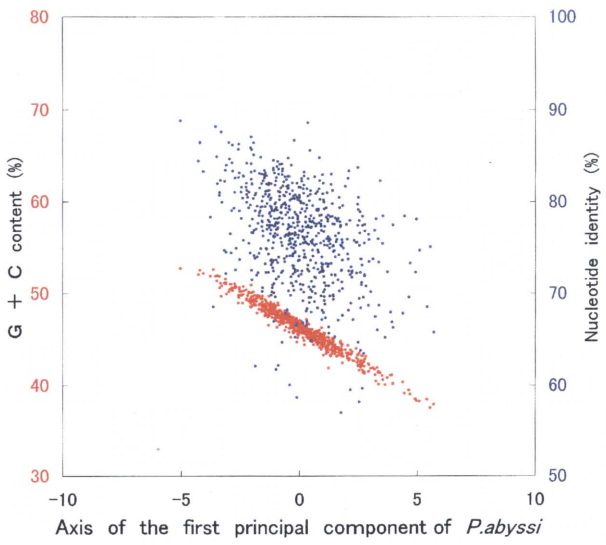

Figure 2. Plotting of genes of *P. abyssi* projected on the first principal component axis against the G+C content of the genes (red dots on the left-hand scale), as well as against the sequence identity between the orthologs (blue dots on the right-hand scale).

Results

1. Distribution of orthologous genes in a trinucleotide composition space

The distribution of the 689 orthologous genes of *P. horikoshii* in the 32-dimensional composition space was projected on the first principal component axis (Fig. 1). The distribution of genes against the G+C content shown as red dots indicates a linear relationship (correlation coefficient 0.98), and that against the degree of sequence identity between the orthologous genes, shown as blue dots shows a weak correlation (correlation coefficient 0.46). The nucleotide sequence identity of 689 orthologous was in the range of 60-90%, and the average G+C content of orthologous was 43.1% for *P.horikoshii* and 46.5% for *P.abyssi*. The same plot for *P. abyssi* projected on the first principal component axis is shown in Figure 2. The distribution against G+C content in red dots is almost a straight line. The distribution against the degree of sequence identity between orthologs in blue dots showed a weak correlation (correlation coefficient -0.44). The sequence identity between orthologs increases with their G+C content both in *P. horikoshii* (Fig. 1) and *P. abyssi* (Fig. 2).

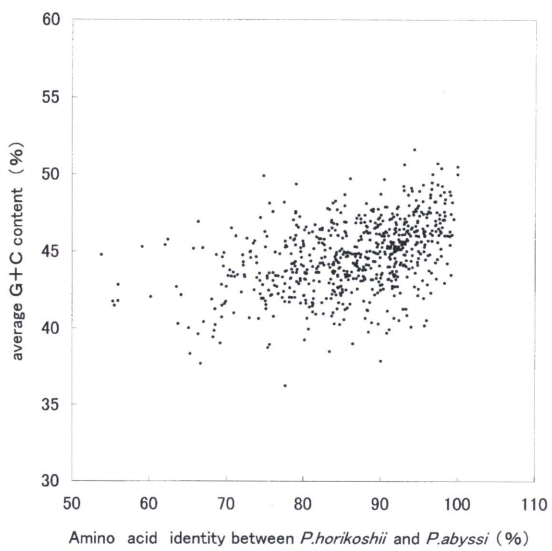

Figure 3. Relationship between orthologous amino acid sequence identity and the average G+C content for pairs of orthologous genes.

The relationship between sequence identity and G+C content was replotted in Figure 3, where the sequence identity is shown in the amino acid scale and the G+C content represents the average of pairs of orthologs. The degree of sequence identity of

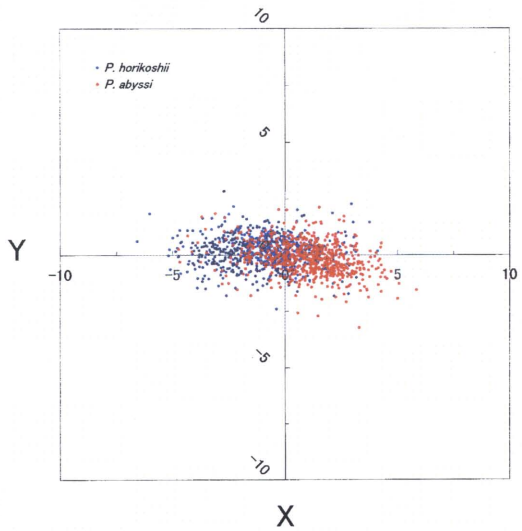

Figure 4. Visual representation of 689 orthologous genes of *P. abyssi* (red dots) and those of *P. horikoshii* (blue dots) in the 32-dimensional trinucleotide composition space projected onto a plane. The plane was defined by the two vectors expressing the first principal component axis for *P. abyssi* and that for *P. horikoshii*.

orthologous amino acid sequences was in the range of 55-100%. The amino acid sequences with a high G+C content showed high sequence identity with a correlation coefficient of 0.41.

The distribution of 689 pairs of orthologs from *P. horikoshii* (blue dots) and *P. abyssi* (red dots), is shown in Figure 4. The x-axis was taken as the first principal component axis for *P. horikoshii*, and the y-axis was taken to be perpendicular to the x-axis on the plane defined by two vectors, one expressing the first principal component axis for *P. horikoshii* and the other that for *P. abyssi*. The origin was the average composition of all genes of the two species. The two first principal component axes are at an angle of 10 degrees, and the gene distribution from the two species indicates a significant overlap. All orthologous genes of both species were classified according to distance. 76% (525/689) of the *P. horikoshii* genes were located close to its average point, and 72% (494/689) of the *P. abyssi* genes were close to its average. This indicates that the orthologous genes are considerably separated in a 32-dimensional composition space than is suggested by the overlap seen in the projection on the plane. Normalized nucleotide compositions with average and standard deviation (equation 2) were used to classify the genes as they showed better discrimination. In Figures 1 and 2, real nucleotide compositions were used as they showed distribution more clearly than did normalized compositions.

2. Nucleotide substitutions between orthologs

Deviations of trinucleotide compositions between two species must be reflected in their nucleotide substitution. The percentage of 155,807 mononucleotide substitutions between 689 pairs of orthologs is shown in Table 1. Adenine (A) in *P. horikoshii* is at 18.8% the most frequently substituted for guanine (G) in *P. abyssi*. Thymine (T) in *P. horikoshii* at 16.2% is the second most frequently substituted for cytosine (C) in *P. abyssi*. Substitutions from A/T of *P. horikoshii* to G/C of *P. abyssi* amount to 46.4% of all substitutions between the two species, while those from G/C of *P. horikoshii* to A/T of *P. abyssi* account for 32.4%. These differences in the substitution rates can be attributed to the difference in the average G+C content,

Table 1. Percentage of mononucleotide substitutions (155,807 in total) between pairs of orthologous genes

	P. abssi				
P. horikoshii	A	T	G	C	sum
A	-	7.2	18.8	6.6	32.6
T	7.2	-	4.8	16.2	28.2
G	13.2	3.6	-	3.4	20.2
C	5.4	10.2	3.4	-	19.0
sum	25.8	21.0	27.0	26.2	100.0

Table 2. Percentage of nucleotide substitutions (90,546 in total) at the third position of synonymous codons

	P. abssi				
P. horikoshii	A	T	G	C	sum
A	-	7.7	18.4	6.1	32.2
T	12.6	-	4.1	15.1	31.8
G	11.5	3.2	-	2.7	17.4
C	5.0	10.9	2.7	-	18.6
sum	29.1	21.8	25.2	23.9	100.0

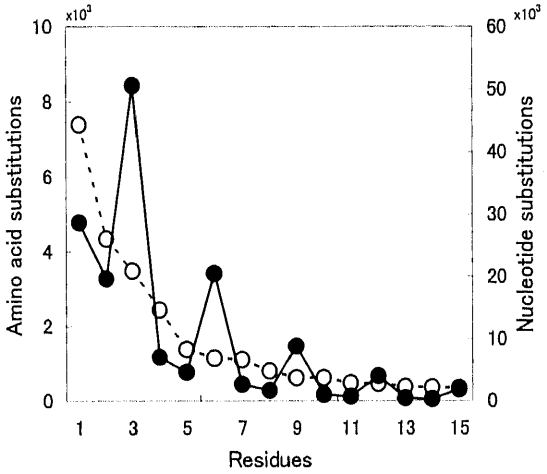

Figure 5. The occurrence of amino acid substitutions against the intervals is plotted in broken line on the left-hand scale. Filled circles indicate the same plot for nucleotide substitutions.

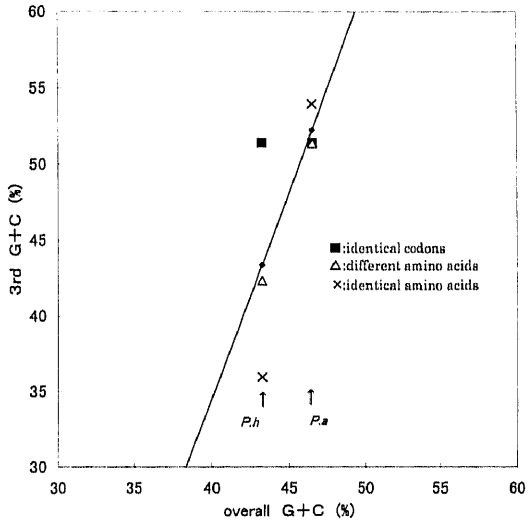

Figure 6. Plotting of the G+C content at the third codon position versus the entire G+C content in lozenges. The G+C content at the third position of IA, DA and IC codons are marked as a cross, a triangle and a square, respectively.

which is 43.3% for *P. horikoshi* and 46.5% for *P. abyssi*. Substitution from A/T to T/A, or from G/C to C/G, between the two species does not change the amount of the G+C content. Substitution from A/T to T/A (14.4%) was observed more than twice as often as that from G/C to C/G (6.8%). This result indicates that G/C is more conservative than A/T, and this is consistent with the correlation between G+C content and sequence conservation. The same tendency was observed for the 90,546 substitutions at the third position of synonymous codons (Table 2).

3. Occurrence of substitutions along the sequence

The occurrences of both amino acid and nucleotide substitutions are plotted against intervals in Figure 5. The interval represents the shortest distance between substitutions along the sequence. The occurrence of substitutions decreased as the interval increased both in proteins and genes. Peaks were observed for every three residues in genes. These peaks may be attributed to the substitutions at the third position of synonymous codons, as they account for 58% (90,546/155,807) of all substitutions.

4. G+C content at the third codon position

Muto and Osawa have reported that genomic G+C content correlates with the G+C content at the third codon position[8], and a linear relationship has been confirmed with the data of many species[7]. Two points representing the total G+C content and G+C

content at the third codon positions of two pyrococcus species are plotted in lozenges and connected by a line in Figure 6. The G+C content at the third position for the three groups of codons has also been marked in Figure 6 (a cross for IA codons, a triangle for DA codons and a square for IC codons). The difference in G+C content at the third position between the IA and IC codons was −15.4 % for *P. horikoshii* and 2.6% for *P. abyssi*. The G+C content at the third position of IC codons of two species is the same according to the definition. The two pyrococcus species have evolved from a common ancestor, and the IC codons are thought to represent their ancestor type. This result indicates that the G+C content of *P. horikoshii* has changed significantly resulting in a reduced G+C content, while the G+C content of *P. abyssi* has remained unchanged.

Discussion

The study presented here has demonstrated that analyzing gene distribution in a composition space is useful not only for classifying but also for detecting directional distribution. The directional distribution of genes must be the result of a certain substitution pattern between the two species. The substitution from A in *P. horikoshii* to G in *P. abyssi* was most frequently observed, and that from T to C was the second most frequent. A substitution from A to G on one DNA strand results from either a direct A-to-G mutation on that strand or a T-to-C mutation on the complementary strand. Therefore, the substitution from A to G and from T to C is actually the same as the change from an A • T pair to a G • C pair. The pattern of all substitutions between the two species (Table 1) was similar to that of substitutions at the third position of synonymous codons (Table 2). This result suggests that the pattern of substitution is independent of the position in the codons. Nucleotide substitutions in a gene sometimes accompany amino acid changes. Substitutions at the third position of synonymous codons yield no amino acid changes and are free of functional constraint, so that they are frequently observed and account for 58% of all substitutions.

The G+C content at the third position of IA codons was 36.0% for *P. horikoshii* and 54.0% for *P. abyssi*, respectively, and that of IC codons was 51.4%. The IC codons were assumed to be the ancestor types of the two pyrococcus species. It can therefore be concluded that the genes of *P. abyssi* are very similar to its ancestor type or have remained unchanged, whereas the genes of *P. horikoshii* have changed resulting in a reduced G+C content. This substitution pattern between two species would be the cause of the directional distribution of orthologous genes in a composition space.

REFERENCES

1) Karlin, S., Burge, C. : Dinucleotide relative abundance extremes : a genomic signature. Trends Genet.,11 : 283-290, 1995.

2) Karlin, S., Mrázek, J. : Compositional differences within and between eukaryotic genomes. Proc. Natl. Acad. Sci. USA, 94 : 10227-10232, 1997.

3) Karlin, S., Mrázek, J., Campbell, A.M.: Compositional biases of bacterial genomes and evolutionary implications. J. Bacteriol., 179 : 3899-3913, 1997.

4) Nakashima, H., Ota, M., Nishikawa, K., Ooi, T. : Genes from nine genomes are separated into their organisms in the dinucleotide composition space. DNA Res., 5 : 251-259, 1998.

5) Nakashima, H., Yamashita, S., Nishikawa, K. : Directional gene distributions of two Mycoplasma species exhibited in the nucleotide composition space. Res. Commun. Biochem. Cell Mol. Biol., 5 : 27-35, 2001.

6) Nakashima, H., Nishikawa, K. : The genomic DNA sequences of various species are distinctively distributed in nucleotide composition space. Res. Commun. Biochem. Cell Mol. Biol., 4 : 25-45, 2000.

7) Bellgard, M.I., Gojobori, T. : Significant differences between the G+C content of synonymous codons in orthologous genes and the genomic G+C content. Gene, 238 : 33-37, 1999.

8) Muto, A., Osawa, S. : The guanine and cytosine content of genomic DNA and bacterial evolution. Proc. Natl. Acad. Sci. USA, 84 : 166-169, 1987.

古細菌ピロコッカス近縁2種間のオルソロガス遺伝子の解析

堀田　宏，中島　広志

要　旨

古細菌ピロコッカス *Pyrococcus horikoshii* と *Pyrococcus abyssi* のオルソロガス遺伝子を3塩基組成，塩基置換頻度，コドン第3位のG＋C含量から解析をおこなった。オルソロガス遺伝子を3塩基組成32成分からなるベクトルとして組成空間に1点で表したとき，分布の第一主成分軸はG＋C含量およびオルソロガス遺伝子間の配列の一致割合と相関を示した。このことから2種の生物種が共通の先祖から分離進化の過程でG＋C含量が変化する方向に進んできたと思われた。塩基AとG，TとCの間の置換頻度が高く，これらはG＋C含量を変化させた。コドン第3位のこれらの置換は同義語コドンに対応し，置換全体の58%を占めた。同一コドン第3位のG＋C含量が2生物種の共通先祖に近いと仮定すれば，*P. horikoshii* はG＋C含量減少の方向に大きく変化してきたが，*P. abyssi* はほとんど変化しなかったと考えられる。