# Chapter 2

# Materials and Methods

## 2.1 Definition of orthologous genes

Orthologous genes are those, having the same function but existing in different species. Therefore, it is reasonable to suggest that the amino acids sequences coded by orthologous genes would show high homology and nearly similar lengths. Thus, the expectation value obtained from the FASTA homology search (E value) [15] indicates the probability of the error in the FASTA homology search results, and the ratio of the overlap length between them. To compare the pairwise alignment, pairs comprising one gene from species 1 and one gene from species 2 were chosen. Then orthologous genes were grouped into the cluster. Additionally, to compare each pairwise alignment, the clusters consisting of more than three genes were removed. Each pairwise alignment was performed by using the CLUSTAL W program, and the pattern of the homology between them was confirmed.

## 2.2 The G+C content and the G+C content at the third codon position

The twenty-three genomic DNA sequences of Archaea and Bacteria used in this analysis were downloaded from the NCBI web site (Table1) [38] .

Both the G+C content of the genomic DNA sequences and the G+C content at each codon position for all ORFs of each species were calculated. All ORFs of *Pyrococcus horikoshii* OT3 (*Ph*) and *Pyrococcus abyssi* (*Pa*) were translated into amino acid sequences. A homology search was done

Table1 Complete genomic DNA sequences used in this analysis.

| No. | Organism | Domain* | Size (Mbp) | Abbreviation | References |
|---|---|---|---|---|---|
| 1 | *Aeropyrum pernix* | A | 1.67 | Aero | 16) |
| 2 | *Archaeoglobus fulgidus* | A | 2.18 | Aful | 17) |
| 3 | *Methanobacterium thermoautotrophicum* | A | 1.75 | Mthe | 18) |
| 4 | *Methanococcus jannaschii* | A | 1.66 | Mjan | 19) |
| 5 | *Pyrococcus horikoshii* OT3 | A | 1.80 | Pyro | 20) |
| 6 | *Pyrococcus abyssi* | A | 1.8 | Pabyssi | 21) |
| 7 | *Aquifex aeolicus* | B | 1.50 | Aquae | 22) |
| 8 | *Bacillus subtilis* | B | 4.20 | Bsub | 23) |
| 9 | *Borrelia burgdorferi* | B | 1.44 | Bbur | 24) |
| 10 | *Chlamydia pneumoniae* | B | 1.23 | Cpnue | 25) |
| 11 | *Chlamydia trachomatis* | B | 1.05 | Ctra | 26) |
| 12 | *Deinococcus radiodurans* | B | 3.28 | Dra | 27) |
| 13 | *Escherichia coli* K-12 | B | 4.60 | Ecoli | 28) |
| 14 | *Haemophilus influenzae Rd* | B | 1.83 | Hinf | 1) |
| 15 | *Helicobacter pylori* | B | 1.66 | Hpyl | 29) |
| 16 | *Helicobacter pylori* J99 | B | 1.64 | Hpyl99 | 30) |
| 17 | *Mycoplasma genitalium* | B | 0.58 | Mgen | 31) |
| 18 | *Mycoplasma pneumoniae* | B | 0.81 | Mpneu | 32) |
| 19 | *Rickettsia prowazekii* | B | 1.10 | Rpxx | 33) |
| 20 | *Synechocystis sp.*PCC 6803 | B | 3.57 | Synecho | 34) |
| 21 | *Thermotoga maritima* | B | 1.80 | Tmar | 35) |
| 22 | *Treponema pallidum* | B | 1.14 | Tpal | 36) |
| 23 | *Mycobacterium tuberculosis* | B | 4.40 | Mtub | 37) |

Characters A and B indicate Archaea and Bacteria, respectively.

in the non-redundant database [39] with each translated ORF of *Ph*, using FASTA program. Furthermore, a homology search was done for all the ORFs of *Pa* with each ORF of *Ph* and vice-versa. Pairs of orthologous genes, including one gene from *Ph* and one gene from *Pa*, were chosen. The pairwise alignment of each pair of orthologous genes was done by using CLUSTAL W [40] and then was applied to calculate the G+C content at the third codon position in the orthologous genes of the two *Pyrococcus* species. In addition, all ORFs were classified into three domains (Archaea, Bacteria and Eukarya), to which the homologous genes belong and/or into eight categories by function, which the homologous genes have [41] .

## 2.3 Division of a DNA sequence by using a Markov model

DNA sequences and the three-dimensional structures of flavodoxin (ferredoxin) reductase [42] , heat shock protein (*grp*E) [43] , flavin oxidoreductase [44] , integrase/recombinase *xer*D [45] , endnuclease III [46] , elongation factor Tu (*tuf*B) [47] in *Escherichia coli* were downloaded from Colibri Database [48] and Protein Data Bank [49] , respectively.

The probability of a given segment of the DNA sequence is calculated by the program GeneMark [50] . The calculation is basically performed as follows: Each DNA sequence is broken up into"windows," typically comprising 96 bases. The probability that this window contains coding sequence (given a previously determined model of the coding sequence trained for a particular species) is calculated. The window is then moved over one"step," typically 12 bases, and the coding probability is calculated again. When the entire DNA sequence has been traversed in this manner, the average of the windows spanning the sequence is computed (The basis of the computation is Bayes Rule.). In this analysis, I calculated the probabilities by the GeneMark program with the third order matrix of Class III genes (*Escherichia coli* horizontally transferred genes) produced by Borodovsky M. *et al* [50] . with 48 bases as the window size and three-base steps. The probabilities of DNA sequence were analyzed according the following definitions.

I compared my results with the domain classified by CATH [51] . In addition, FASTA homology search were done to the DDBJ ALL database (DDBJ release + updates Non-redundant nucleotide sequences Database and/or protein Database) [52] with the divided regions of the DNA sequences of flavo-

doxin reductase and heat shock protein (*grp*E) .

In the case that the region is the same class adjacent to the region, the two regions are considered as a region because of being from the same species. About the homology results using DNA sequence as a query, to prevent analysis of wrong sequences and too short DNA sequences, the genes were chosen whose E value (probability of the wrong sequences existing in the search result) is less than 1.0, and ratio of length overlapped between query sequence and subject sequence is more than 0.5. About the homology results using amino acid sequence as a query, E value is less than 0.1. The genes from Eukarya, and genes whose DNA sequence is not a full cDNA but a fragment, were excluded. Also the genes homologous to both DNA sequence of the two domains, were excluded since it is not meaningful for discussion of evolution. Both domains of flavodoxin reductase and heat shock protein (*grp*E) were aligned by using CLUSTAL W program.

## 2.4 Definitions of class, region and CLASS

Médigue C. *et al.* suggested that the codon usage of 708 genes in *Escherichia coli* could be classified into three classes (Classes I, II, and III) by Fractional Correspondence Analysis and the Dynamic Clustering Method. Class I genes, with intermediate codon usage bias, maintain a low or intermediate level of expression, although some genes may occasionally be expressed at a very high level under environmentally rare conditions. Class II genes, having high codon usage bias, are highly expressed under exponential growth conditions. Genes from Class III, with low codon usage bias, included the following genes for fimbriae, flagellae and pili, integration host factors (*hip* and *himA*), genes controlling cell division (*dic*ABC, structurally related to template phages), several outer membrane or periplasmic protein genes and several catabolic operons (threonine degradation, $\beta$-glucoside degradation, fucose degradation), genes containing insertion sequences, and genes behaving as mutators when inactivated (*mut*H, *mut*T, and *mut*D) and lambdoid phage lysogeny control protein [53] . Also they suggested strongly that these Class III genes mostly comprised genes inherited by horizontal transfer. Borodovsky M. *et al.* analyzed the DNA sequence of *Escherichia coli* by means of a Markov model and showed that only the results to be identified as the Class III genes in the genome DNA sequence, using matrix produced by Class III genes, were allowed with acceptable accuracy [54] . Accordingly, to explore the DNA sequence from other species, where structural information had been published, I used the third-order matrix, produced

with Class III genes by Borodovsky M. *et al.* (*Escherichia coli* horizontally transferred genes) for my analysis. The probability indicates the degree of similarity to the DNA sequence of those genes that were used for producing the matrix with the Markov model.

Therefore, when the class of region is high, the probability of the region is high. This means that this DNA sequence is similar to the genes, those were used for making matrix. When the class of region is low, then the probability of the region is low. This means that this DNA sequence has low similarity to the genes, those were used for making matrix. Accordingly, it is supposed that the greater the calculated probability score of the DNA sequence, the more similar is the codon usage of the genes, by using those matrices produced by Markov models. This indicates that there is more similarity between the analyzed sequence and the sequences of the genes used for producing the matrices as score becomes closer to 1.0. When the score is closer to 0.0, there is less similarity between them. In contrast, when the score of the probability is close to 0.5, the degree of similarity among them could not be determined. Therefore, meaningful scores of the probability were considered to be greater than 0.6 or less than 0.3. Then I grouped the probability scores into six classes by steps of 0.1, as given in Table 2.

The probability was divided into regions when classes changed. But I do not take into account the region whose length is shorter than or equal to 48 base-pairs, since the probability is calculated with 48 base-pairs as a window, and the positions of the class 2, since the degree of similarity among them could not be determined.

When the class of probability ranges from class 3 to class 6, the frequency the same class of probability should be more than or equal to five. Also when the frequency of class 1 is from 3 to 12 continuously, then that position was the boundary between two regions. When the frequency of class 1 is more than or equal to 13 continuously, without consideration of positions ofthe class 2, then the continuous positions are regarded as one region.

When the frequency of the same continuous class is more than or equal to five, the CLASS is regarded as a class, by itself. Conversely, when the frequency of the same continuous class is less than six, or if no class occurs continuously, then the most frequent class is taken as the CLASS.

Table 2 Classification of Probability of DNA sequence

| Class | Score of probability of DNA sequence |
|---|---|
| 1 | less than or equal to 0.3 |
| 2 | greater than 0.3 and less than 0.6 |
| 3 | greater than or equal to 0.6 and less than 0.7 |
| 4 | greater than or equal to 0.7 and less than 0.8 |
| 5 | greater than or equal to 0.8 and less than 0.9 |
| 6 | greater than or equal to 0.9 |