Chapter 1

Introduction

Since life began on earth about 4 billion years ago, numerous species, numbering more than a million, are estimated to be living now on our planet. The diversity of the living life-forms is the result of evolution, by which variation has been accumulated. The trace of these variations has been recorded on the DNA sequence. The information in the DNA sequence of a gene is transmitted accurately by the duplication of a cell, from one to another. Evolution is that a wide variety of characters, such as those influenced by genetic changes, have been accumulated over generations, hierarchically structuralized, and ordered in a complicated manner. It is the main role of a gene to convey the genetic information. Therefore, the design of life is recorded on DNA sequences of chromosomes. And this design changes in the process of evolution, by such means as horizontal gene transfer, rearrangement, and fusion of DNA sequences, and mutations. The trace of variations has been recorded on DNA sequences of chromosomes, too. DNA sequences of a gene, on which the variation has been coded, are transcribed into mRNA, and continuously mRNAs are translated into amino-acid sequences. Furthermore, an amino acid sequence becomes a polymer compound, which then forms a three-dimensional structure to express its function as an enzyme. A sequence of three nucleotide residues or codon is required to specify one amino acid residue; however most one codon. Therefore, to reveal the evolutionary process in organisms, it is better to analyze DNA sequences, onto which the variation has been directly recorded, rather than amino acid sequences translated from DNA sequences as an ambiguous template. Unfortunately, until the method to determine genomic DNA sequences was established and the DNA sequences of many species became available, only the amino acid sequences translated into a protein could be analyzed. In 1995, Fleischmann R. D. *et al.* determined the genomic DNA sequence of *Haemophilus influenzae* Rd [1] . In 2002, genomic

DNA sequences of 16 species from Archaea, 72 species from Bacteria, and 6 species from Eukarya were determined and published (see appendix). Approximately 90% of genomic DNA sequences from the Human species were also opened to the public [2, 3] . Genomic DNA sequencing projects of many organisms are now in progress. Therfore, I can use the DNA sequences as a target for directly analyzing the process of evolution, now.

For the analysis of a DNA sequence, the G+C content, G+C content at the third codon position, and the codon adaptation index (CAI) were mainly used. Lawrence J. G. and Ochman H. suggested that, as a minimum, 18% of open reading frames (ORFs) with typical nucleotide composition and codon usage, representing over 600 genes, might have arisen through lateral gene transfer in the complete *Escherichia coli* strain MG1655 [4] . This amount was estimated by analyzing the base composition and codon bias of the complete *E. coli* strain MG1655. Koski L. B. *et al.* reported that when ORFs located in the complete *E. coli* strain MG1655 and *Salmonella typhi* were extracted with the G+C content, G+C content at the third codon position, and the CAI as candidates for transferred genes, these could identify the transferred genes in *E. coli*, by constructing a phylogenetic tree with these candidates [5] . CAI is a measure of similarity of a gene's synonymous codon usage to that of a standard set of highly expressed genes for that organism [6] . They also estimated that 48% of the phylogenetic trees for novel genes, which might have been either deleted from *Salmonella typhi* or introduced into *E. coli*, indicated horizontal transmission. This implied that about 10 to 15% of the total *E. coli* ORFs was the result of introgression [5] . Moreover,they suggested that many genes have been introduced into the *E. coli* genome, and that was one of the driving forces for there being new genes transferred from one species to other. It is very difficult to identify the species, from which the laterally transferred gene might have come, although the ratio of lateral genes to all genes was large. The relationship between the three-dimensional structure of a protein and transferred genes had not been mentioned.

Doolittle R. F. reviewed the multiply origins of micro organisms, especially through gene transfer, by analysis of amino acid sequences that were translated from ORFs in the genomic DNA sequence [7] . Both the degree of similarity between sequences (the new ORF and those in the database), and the closeness in the relationship between the organisms, from which they were obtained, were used to predict the function for the identified ORFs in the genomic sequences. This is a first step to identify ORFs more accurately for prediction of their function. It is also important to identify the function of ORFs to analyze genome characteristics, gene organization, gene transfer or gene fusion, etc.. However, Doolittle noted that"it is not a foolproof method for ascertaining gene function, and it is often complicated by horizontal gene

transfers." Indeed, although many transferred genes have been identified, there are few reports about the three-dimensional structure of protein with respect to the DNA sequence of a transferred gene.

I thought that the process of evolution is led to be frequently misunderstood by enormous gene transfers. Some examples were found by doing comparison among the whole genomic DNA sequences of Bacteria and Archaea. By the comparison between virulent and avirulent strains of *Helicobacterium pylori*, a pathogenicity island, [8] which is a 40,000-bp 'island' of DNA that includes a large number of genes involved in attacking host cells, was found in the virulent strain [8] . Avirulent strains of this bacterium lack this region. *Bacillus anthracis* and a region related to symbiosis, have regions similar to the pathogenicity island of *Helicobacterium pylori*, that is, *Bacillus anthracis* (the cause of anthrax) contains two large plasmids, one of which has a 44.5-kb pathogenicity island [9] . This island, which contains genes for the toxin that can be very lethal to humans, is flanked by inverted insertion sequences. The plasmid also contains a collection of what seem to be transposases and integrases, suggesting a history of shuffling and exchange. Not all genomic islands encode genes for pathogenicity. In *Mesorhizobium loti*, for example, the main chromosome has a 'symbiotic island' of more than 6,000,000 bp, which is necessary for the bacterium to establish a symbiosis with its legume host plant. It is flanked by 17-bp repeats and has a codon usage that is significantly different from the rest of the chromosome [10] . This suggests an 'alien' origin. It is said to be a usual thing, where a region has been deleted or introduced from other species. This means that many genes might be transferred for a new function as a cluster of genes, not for creating a functional gene. Though it might be a good way to gain a new gene, the structure of the coded protein was not mentioned.

By the analysis of DNA sequences, it is suggested that only transferred genes have been occurring, not for creating but for gaining new genes. The structure of the protein coded for by a transferred gene had not been mentioned at all. Thus, I thought that it is not always true that all genes might have transferred from other species, in order that new proteins could originate in the evolution of life. As an example, it is a possibility that a protein with a new function could be produced by fusion between its own genes or between its own gene and a gene, which might be transferred from another species. The fusion of such genes is a fundamental design to create a new function. Moreover, a new protein would have been created by the fusion, which would form complex domains (multi-domains) designed by motifs of amino acid sequences. This is one of the essential ways to increase the complexity of function in a new protein.

Fis (factor for inversion stimulation), a recombinational enhancer in *E.*

coli, is a protein of multi-functions, such as the adjustment of gene expression, *ori*-directed initiator for chromosome replication, and many site-specific recombinations. Morett E. and Bork B. suggested that Fis protein in the α-proteobacterial species forms an operon with *ntr*C and *nifR*3, which belongs to a NifR3 family, one family of TIM-barrel proteins in PFAM protein families. Fis is a putative fusion protein consisting of an amino acid sequence at the C-terminus from *ntr*C and that of *ydh*G, a *nifR*3 homolog, by analysis of the phylogenetic tree with amino acid sequences [11] .

To search for fused genes in the genomic DNA sequence, Koonin E. V. *et al.* constructed a phylogenetic tree with amino acid sequences translated from genes and a cluster of those of proteins of similar function in different species (COGs) [12] , and then analyzed domains (Archaea, Bacteria, and Eukarya) derived from those species, to which the ORFs or proteins belong [13] . As a result, 51 fused genes out of 405 identified fused genes were demonstrated to have belonged to at least two domains. Furthermore, 31 fused genes out of 51 might have been horizontally transferred from one domain to another, and 14 fused genes out of 51 might have been evolved only in its own domain. Thus two scenarios for the process of evolution were found: one is the transfer of a gene horizontally from one domain to another domain, and the other is the gene transfer between the same domains. But the occurrence of gene fusion independently within its domain seems to be rarer than horizontal gene transfer [13] .

Amino acid sequences of ORFs show a correlation with the structure of protein. But to identify where a gene has been transferred or fused, the coded proteins are needed to construct a phylogenetic tree. If there are no homologous proteins in the database, the gene cannot be identified. Thus, it is not sufficient to analyze only the DNA sequence for identifying transferred genes by means of the G+C content, etc., or only amino acid sequences for finding the transferred genes or fused genes by constructing a phylogenetic tree. Since the trace of variations has been recorded on the DNA sequence directly, it is important to develop a method for demonstrating the relationship between the DNA sequence and the structure of the protein from a gene by directly analyzing DNA sequence. Therefore, I classified the events of gene transfer or gene fusion into two types, which is supposed to have created or gained a new protein function. One is a lateral gene transfer from another species. The other is a fusion between genes in the chromosome or between a lateral gene and its own gene. Amino acid sequences coded by genes are generally used for identification of lateral genes and/or fused genes, since these events are usually analyzed on the basis of a phylogenetic tree. However, differences in the G+C content in some regions of the chromosome were pointed out for demonstrating the presence of lateral genes or a region

from another species. Since variations have been recorded directly on DNA sequences, it is much better to analyze DNA sequences rather than amino acid sequences for finding essential information. Thus it is an important key step for gaining a better understanding of evolution that a putative horizontally transferred gene or a putative fused gene are identified in the genomic DNA sequence.

Accordingly, I thought that the following two methods would be useful to identify a putative horizontal transferred gene or a putative fused gene in the genomic DNA sequence. Generally, the G+C content, the G+C content at the third codon position, and codon usage are used to analyze and identify horizontal gene transfer from other species. Koski L.B. *et al.* identified genes transferred from *Salmonella typhi* to *E. coli* by the G+C content, the G+C content of synonymous codons at the first or the third codon position, etc. They defined the orthologs as follows: the single most similar *S. typhi* ORF for an *E. coli* ORF was obtained by BLASTN [14] with a cutoff of 10.0, and the alignment lengths were used to further evaluate the similarity between the genes. But even if the *S. typhi* ORF showed the most similarity to an *E. coli* ORF, it does not always hold that the *E. coli* ORF shows the most similarity to the *S. typhi* ORF. It is important for the two genes to show homology to each other, since orthologous genes are suggested to have existed before the division of the two species. Also, they did not give much consideration to the structure. Thus, one method that I applied was to analyze the G+C content at the third codon position for identifying a putative lateral gene of similar function in different species and of conserved protein structure among the orthologous genes, especially in the Archaeal domain. The genomic DNA sequence of two very closely related *Pyrococcus* species, *Pyrococcus horikoshii* OT3 and *Pyrococcus abyssi*, had been determined and found to be very small. The other thing I found is that it is important to search for a statistical difference in the DNA sequence of a gene to find a fused gene. To do so, I used a Markov model for this analysis, since the probability of DNA sequence calculated by a Markov model indicates similarity to the DNA sequences of the genes that were used to produce the matrix for calculation of probability. Moreover, it can be said that the higher the probability is, the more similar the genes is to the character of the genes used to produce the matrix. Thus I thought that the region divided by the probabilities calculated by the Markov models might represent the difference of origin.

I considered that the structure of a protein has to be taken into account when the protein for a gene expresses its function as an enzyme. My aim is to correlate the process of evolution by such events as horizontal gene transfer or gene fusion, which might have occurred, with information on protein structure.

I hope I can clarify the process of protein evolution from the point of view of the *information* of protein structure.