Chapter 6

Discussion

Since the begining of life on earth about 4 billion years ago, the numbers of life-forms have increased, with an estimatal of around more than a million species co-habiting on the earth, now. The diversity of the living beings is the result of evolution, by which more and more variations have been accumulated on the DNA of a gene. The variation is accurately transmitted by the duplication of a cell from one generation to another, and has been recorded directly onto the DNA of chromosomes. These variations in evolution can be demonstrated by the direct analysis of DNA sequences. Genetic Transfer is the main role of a gene. Therefore, the blueprint of a life-form is recorded on the DNA sequence of chromosomes together with the changes occurring in the process of evolution, such as horizontal gene transfer, rearrangement, fusion of DNA sequences, and mutation. DNA sequences of a gene, onto which variations were recorded, are transcribed into mRNA, and, continuously, mRNAs are translated into amino acids sequences. Furthermore, an amino acid sequence becomes a polymer compound, and then forms a three-dimension structure to express its function as an enzyme. An amino acid can correspond to not only one codon in the process of transcription. In the past, long before the method to determine genomic DNA sequences was established, only amino acid sequences of a protein could be analyzed. Many genomic DNA sequences are now available. By directly analyzing the DNA sequences, upon which changes have been directly recorded, rather than amino acid sequences, which are translated from the DNA sequences as ambiguous templates, we are able to find more precise results of the process of evolution.

In 1987, when it was difficult to analyze the DNA sequence, Watson J. D. *et al.* proposed that proteins can be classified into three types, by comparison with those from Humans [85] . The "Ancient Proteins" were classified into

two subgroups, "First editions" and "Second editions." Proteins belonging to "First editions" had functions related to the main reactions of metabolism, which have not changed since branching between Bacteria and Eukarya. An example from among these proteins is triose phosphate isomerase, whose score of identity is 46% between the amino acid sequence from Humans and that from Bacteria. Basic processes, such as replication of DNA sequence, transcription, metabolism, and protein synthesis are remarkably similar in all living organisms. One important thing about this situation is that structural changes in the proteins from house-keeping genes cannot be allowed, because their function is essential for life. The proteins belonging to "Second editions," which are common proteins in both Humans and Bacteria, show similar identities with each other with respect to amino acid sequence, but have different functions. An example is the glutathione reductase in blood cells of Humans and mercury reductase in the bacterium *Pseudomonas*. The score of identity of amino acids between these two proteins is 27%.

The "Middle-age Proteins" consist of proteins that are found in most Eukarya, but for which counterparts in Bacteria are as yet unknown. An example is actin.

The "Modern Proteins" are proteins composed of the following three subclasses: (1) "Recent vintage" are proteins found in animals or plants, but not both, and are not found in Bacteria; (2) "Very recent inventions" are proteins found only in vertebrates, but not elsewhere; (3) "Recent mosaic" are modern proteins clearly produced by shuffling of exons. Collagen, plasma albumin, and low density lipoprotein receptor are examples of proteins belonging to the subclasses (1), (2), and (3), respectively.

Thus, in the process of evolution, many proteins have not changed essentially (corresponding to the proteins in "First editions"), but some have come to have a different function by the accumulation of changes (corresponding to the proteins in "Second editions"). Additionally, some proteins have been produced in an evolutionally later time (corresponding to the proteins in "Middle-age Proteins" or the kind of proteins in "Modern Proteins"), and some have been produced by a combination of fragments, which existed in other proteins. As a result of this comparison, two things became clear. One is that during evolution, many genes did not need to change, and if needed, the change was very slight. The other is that some new proteins might have been produced by the incorporation of DNA sequences or fusion between them, which is based on the findings that proteins would have also been produced by the combination of fragments of the existing proteins.

I thought that the same process is likely to have also occurred in proteins of Bacteria. The three-dimensional structures of some functional proteins, which were determined by X-ray crystallography, were composed of

multi-domains. Multi-domains are considered to be constructed by several domains. For some of the functional proteins, it is possible that each part incorporated into the existing structure might have been coded for by a gene from another species. In that way, structurally divided multi-domains might have been composed by those parts of proteins to create a new function during evolution. Furthermore, the protein structures coming from those genes might have been produced not only by originating from the gene by itself, but also by being composed of several parts of a structure from other species, through such events as lateral gene transfer, fusion of DNA sequences and rearrangement of DNA sequence, etc., resulting in the creation of a new function. To demonstrate these events, it is important to extract the information of variation from DNA sequences. Firstly, I applied the G+C content at the third codon position of orthologous genes in the genomic DNA sequence to find out the lateral genes among the genes in which variations have been unaffected or slightly influenced. Next, I analyzed the DNA sequence of a gene with a Markov model to search for genes coding for proteins in which each part of the existing structure of the protein coded for by a gene from other species might have been incorporated; then structurally divided multi-domains might have been composed of those parts, and in that way, created a new function.

I analyzed 1217 orthologous genes of the genomic DNA sequence of *Pyrococcus horikoshii* OT3 by using the G+C content of synonymous codons at the third codon position to identify lateral genes. Orthologous genes, which exist in different species and show the same function, are suggested to be genes in an organism before it underwent branching in evolution. By analysis of the G+C content of synonymous codons at the third codon position, however, I found that the region from 300 kbp to 420 kbp showed a high G+C content of synonymous codons at the third codon position with a big bias compared with that of its own orthologous gene. Twelve orthologous genes, corresponding to approximately 75% of the orthologous genes in the region, thus, came from Bacteria or Archaea. Further, 4 orthologous genes, corresponding to approximately 30% of the orthologous gene in the region, were homologous to genes of function related either to the cell envelope or to biosynthesis of surface polysaccharides and lipopolysaccharides from *Streptococcus*. In contrast, the organization of the operon related to this function is known to differ among *Streptococcus thermophilus* species. 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

G+C content in this region showed variability. This suggests that the entire *epsL-IS981SC* region was transferred from *L. lactis*. Thus, these genes were reported to be putative horizontally transferred genes. 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 8 putative lateral genes were located in this region. Additionally, 15 putative lateral genes were identified in the high GC3 content region. Thus, it is clear that, even if genes would be considered orthologous, they do not always exist in an organism before branching of two species. This means that there is a possibility that such orthologous genes could lead us to misunderstand the phylogenetic relationship during the evolutionary process. Moreover, I suggested that a lateral orthologous gene might actually be a horizontally transferred gene, that is having a different DNA sequence but similar amino acid sequence and structural homology.

Next, I did a gene search using a Markov model, such that the DNA sequences would be incorporated as parts of the gene in other species to create a new function by consisting of a new domain structurally separated. I analyzed 6 genes in *Escherichia coli*, flavodoxin reductase, flavin oxidoreductase, integrase/recombinase *xer*D, endonuclease III, heat shock protein (*grp*E), and elongation factor Tu (*tuf*B), by using the Markov model produced by Borodovsky M. *et al.* with the Class III genes (*Escherichia coli* horizontally transferred genes) in *Escherichia coli*. Five of the 6 genes, with the exception of flavin oxidoreductase, were demonstrated to be divided in correspondence with the domains as parts of the structure by using the probability of the DNA sequence. Especially, flavodoxin reductase has two domains, the FAD domain and the NADP domain, which are related to function. The divided regions from I to III belonged to the FAD domain, where FAD is bound, and those from IV to VIII belonged to the NADP domain, where NADP/NADPH is bound. These two domains exemplified the correspondence to the domains classified on the basis of "Class," derived from the secondary structure, "Architecture," derived from the gross orientation of the secondary structure, and "Topology" defined by CATH. Moreover, the divided regions corresponded to the domains defined by CATH and those determined by X-ray crystallography and the gross secondary structures. I showed that the regions divided by the probability calculated with a Markov model showed the degree of similarity to the DNA sequences from lateral genes (horizontally transferred genes in *Escherichia coli*) and that structural information would have been recorded on the divided regions by the probability calculated with a Markov model. Only region III of flavin oxidoreductase was bridged between two domains. Looking into more detailed structure, I could include one helix into the former domain among the two domains

classified by CATH. I also noted that certain genes exist, such as that for flavin oxidoreductase, that were not incorporated from the DNA sequence as a part of structure. Although there are genes that are not divided into regions corresponding to structure, I was able to find the genes that are divided into regions corresponding to structure by this method. Accordingly, it can be said that the methodology of the Markov model is a useful tool for dividing the DNA sequence of a gene into regions corresponding to the coded protein structure.

Subsequently, I analyzed two genes in more details, flavodoxin reductase and heat shock protein (*grp*E) in *Escherichia coli*. At first, it was demonstrated that the DNA sequence of flavodoxin reductase was divided into seven regions as shown in Fig. 16, and the divided regions also showed a correspondence to the domains, which show correspondence to those defined by CATH, of the three-dimensional structure of flavodoxin reductase determined by X-ray crystallography. Moreover, they correspond to the gross secondary structures of the protein as shown in Fig. 17. Based on the results from homology searches with these sequences, the DNA sequence of the region I+II might have been derived from the N-terminus of *Azotobacter vinelanndii* NADPH:ferredoxin reductase, in which the FAD binding site is located. On the level of the divided DNA sequence, the DNA sequence of the region I+II at the N-terminus of the flavodoxin reductase, showed high homology with an E-value of 0.0028 and identity of 55.80%, to that sequence at the N-terminus of *Azotobacter vinelanndii* NADPH:ferredoxin reductase in my analysis. On the level of amino acid sequence, the amino acid sequence of the FAD domain showed similarity with 44.0% identity to that at the N-terminus region and with 32.3% identity to that at the C-terminus region of NADPH:ferredoxin from *Azotobacter vinelanndii*, respectively. The three-dimensional structures of flavodoxin reductase from *Escherichia coli* and NADPH:ferredoxin reductase from *Azotobacter vinelanndii* were determined by X-ray crystallography. And the RMSD of the aligned C_α atoms of the homologous amino acid residues between NADPH:ferredoxin reductase in *Azotobacter vinelanndii* and flavodoxin reductase of *Escherichia coli* was 2.01Å. This means that the overall structure between them is very similar. The DNA sequences of the divided regions I-III for Flavodoxin reductase possess information in correspondence with the three-dimensional structure and that of the divided region I+II was homologous to the DNA sequence at the N-terminus for NADPH:ferredoxin reductase in *Azotobacter vinelanndii*. These results also suggested that there exists a relationship among the three-dimensional structure, amino acid sequence, and DNA sequence corresponding to the domains of the three-dimensional structure, between flavodoxin reductase from *Escherichia coli* and NADPH:ferredoxin reductase from

Azotobacter vinelanndii.Thus, flavodoxin reductase from *Escherichia coli* and NADPH:ferredoxin reductase from *Azotobacter vinelanndii* would have been incorporated into their respective genomes without incurring change in the three-dimensional structure or domains of the three-dimensional structure during evolution. As a result, differences between the DNA sequences of the gene from *Escherichia coli* and that from *Azotobacter vinelanndii* might have been increased.

The three genes, which the DNA sequence of NADP/NADPH domain is elucidated to have been derived from, have the following characters; *Ralstonia* sp. CH34 pMOL30 *czc*B (membrane fusion protein) is one of three genes conferring resistance to zinc, cadmium, and cobalt. The *Escherichia coli* pKM101 *tra*E gene (conjugal transfer gene E) is located on the plasmid with the deleted region of heavy metal resistance and drug resistance from IncN plasmid R46. *Streptococcus pneumoniae* bacteriophage Cp-1 orf17 is thought to function as a tail protein. *Salmonella typhimurium* phage P22 tailspike protein could be inserted with the viral DNA sequence at the N- and C-terminus regions and these mutants could be expressed at high levels in the host cells and function properly [86] . This suggests that the DNA sequence of phage tail protein and DNA fragments from other species might be easily fused together when they become incorporated into a cell. Thus, the regions I+II, IV+V, VI+VII and VIII of flavodoxin reductase in *Escherichia coli* would have been derived from NADPH:ferredoxin reductase from *Azotobacter vinelanndii*, conjugal transfer gene E (*tra*E) from *Escherichia coli* pKM101, *czc*B (membrane fusion protein) from *Ralstonia* sp. CH34 pMOL30, and orf17 from *Streptococcus pneumoniae* bacteriophage Cp-1, respectively.

I suppose that the DNA sequence of flavodoxin reductase from *Escherichia coli* shows homology to that of the FAD domain of NADPH:ferredoxin reductase encoded on the chromosome of *Azotobacter vinelanndii*, and that the DNA sequence of the NADP domain from *Escherichia coli* is comprised of those sequences of genes on plasmids or bacteriophages. The high probabilities of the NADP domain corresponded to the DNA sequences of genes from other species. The gene encoded on a plasmid or phage and the gene encoded on the chromosome are located in distinct components within the cell. Therefore, recombination of NADPH:ferredoxin reductase from *Azotobacter vinelanndii* might have ocurred since the DNA sequences might have been incorporated the gene related to heavy metal resistance encoded on a plasmid, and the gene of function related to tail protein, encoded on a bacteriophage (horizontally transferred gene) and then the two domains, the FAD and the NADP domains, would have been fused. As a result, the structure and function of the NADH-binding NADPH flavodoxin reductase would be acquired in the process of evolution. It can be said that the results from a

homology search with these sequences, such as genes, species, and identity, also contain information concerning the corresponding with the domains of the three-dimensional structure. Accordingly, I suggested that the regions for flavodoxin reductase divided by the Markov model might have resulted as an incorporation of the DNA sequences from other species, where structural information as parts of the whole structure had been recorded. Thus, I presented a hypothesis that the structure of a gene might have been produced not only by the originating gene by itself, but as a composite of several parts of structure from other species, through events such as gene transfer, fusion of DNA sequences, and rearrangement of DNA sequences, etc. and have created a new function in evolution, by analysis of the functional gene (enzyme) flavodoxin reductase in *Escherichia coli* using a Markov model.

It is also clear that there is a difference between the amino acid sequences and the DNA sequences of flavodoxin reductase from *Escherichia coli* and those of NADPH:ferredoxin from *Azotobacter vinelanndii*, but the three-dimensional structures of these proteins from both species are conserved. By the analysis of the G+C content at the third codon position of synonymous codons, I suggested that there might be genes in *Pyrococcus horikoshii* OT3 in which DNA sequences would have been changed but the three-dimensional structures would not be affected during evolution [87] . In this case, there are differences between the DNA sequences of the orthologous genes, but the structures and amino acid sequences of the proteins from them are estimated to have been conserved. Thus, it is thought that there are many cases of similar relationships between the three-dimensional structure, or the domains of the three-dimensional structure, and the DNA sequence of a gene, and this represents the information accumlated in evolution, which has been recorded on the DNA sequence.

Next, I analyzed in more details the DNA sequence of heat shock protein (*grp*E) in *Escherichia coli*. The DNA sequence of heat shock protein (*grp*E) could be divided into six regions, from I to VI. The three-quarters from the end of region II, and the region III belonged to the domain at the N-terminus determined by X-ray crystallography, the class and architecture of which, in accordance with the definition of CATH, are "Alpha Beta" and "Complex," respectively. The regions from IV to VI belonged to the domain at the C-terminus, the class and architecture of which are "Mainly Beta" and "Roll," respectively.

Homology search was also done within the DDBJALL database (nucleotide sequence) with the regions I+II+III, IV+V, and VI. The genes homologous to the DNA sequences of the three regions werte excluded, since it is not meaningful for the discussion of evolution. No homology was shown to the region I+II+III found in the DDBJALL database (nucleotide sequence).

Region IV+V was homologous to the DNA sequence for the transcription antitermination protein (*nus*G) from *Streptomyces coelicolor*. Region VI showed homology to the DNA sequence for heat shock protein (*grp*E) of *Nitrosomonas europaea*, or that for a very large tegument protein (UL36) of *Gallid herpesvirus* 1 (serotype 2). These results do not seem to make a good scenario of having incorporated DNA sequence as parts of structure.

By looking at the pattern of class of probability, the pattern of the region from I and the front half of region II was similar to that of the region III. Homology search was also done within the DDBJALL database (nucleotide sequence) for the two regions that correspond with the region I+II and the region III-VI. In the contrast to the above findings, the same genes with high similarity were chosen. The two regions were homologous to almost the same region of the *grp*E from the same Bacterial species, but not from any Eukarya species or any Archaeal species. The regions, which were homologous to the regions I+II and the region III-VI, were overlapped. By analysis of the alignment of the two regions of the region I+II and the region III-VI, the nucleotide identity was approximately 47%. Therefore it seems that the *grp*E gene consists of the fusion of the two fragments after duplication of the DNA sequence. Moreover, the secondary structures of the two domains defined by CATH were composed of an α-helix and β-sheet, respectively. I suggest the possibility that a frame-shift might have occurred, because of deletion and/or insertion in the case of the duplication of the fragment. In this way, the DNA sequences would be translated into different amino acid sequences in comparison with those translated from the original DNA sequence. The frame-shift supports that the different amino acid sequence might have formed different secondary structures.

Subsequently, FASTA homology search in the DDBJ protein database was done with the amino acid sequence of GrpE in *Escherichia coli*. Homologous genes were obtained as follows. GrpE from 7 species, *Bacillus subtilis*, *Geobacillus stearothermophilus*, *Yersinia pestis*, *Clostridium acetobutylicum*, *Streptococcus pyogenes*, *Thermus thermophilus*, and *Myxococcus xanthus*, of Bacteria; from 1 species of Archaea, *Methanosarcina mazei*; and the GrpE homolog from *Saccharomyces cerevisiae* mitochondria, of Eukarya, as shown in Fig. 22. By analysis of the alignment of GrpE from 10 species, amino acid sequences at the C-terminus were found to be more well-conserved than those at the N-terminus. This result was similar to that reported by Osipiuk J. and Joachimiak A [88] . The GrpE protein from an archaeon *Methanosarcina mazei*, was suggested to be diverged before branching to Bacteria and Archaea [89] . Two hybrid systems of the bacterial chaperone machine and eukaryotic chaperonin system function in the same cell of *Methanosarcina mazei*. The protein product was the bacterial homologs.

They noted: "It was concluded that the GrpE are bacteria-like and, possibly, of bacterial origin. Interestingly, all extreme thermophile Bacteria investigated to have the gene, which is opposite of the situation in extreme thermophilic Archaea [90] ." The MDE1 protein from *Saccharomyces cerevisiae*, which is a homolog of the GrpE from Bacteria, was also included and exists in the mitochondrial organ [91] . The mitochondrial organ is presumed to have been originated from symbiosis between Archaea and Bacteria and is proposed to be of Bacterial origin in the cells of the Eukarya. The MDE1 protein is suggested to be of Bacterial origin. Accordingly, the GrpE protein came to have the present function by fusion after duplication of the region at the C-terminus in the Bacterial domain. And then, the gene before duplication seems to have already disappeared.

In this work, I was able to identify lateral genes, which might have conserved amino acid sequences and the three-dimensional structures, among the orthologous genes by using the G+C content of the synonymous codons at the third codon position. Also using a Markov model, I could identify the gene with a structure that might have been produced by the composition of several parts of a structure from other species, through such events as gene transfer, fusion of DNA sequences, and rearrangement of DNA sequences to create a new function during evolution. In addition, I was able to demonstrate the duplication of a gene, and then change of the secondary structures through duplication in the process of evolution, which was not found by analyzing only amino acid sequences.

My work opened a new view that direct analysis of the genomic information leads us to understand the evolution of proteins. By analysis of DNA sequences, which has recorded all information of the living life-forms, and progress of evolutionary changes, I can clarify the process of evolution.