Chem-Bio Informatics Journal, Vol. 2, No. 4, pp.119-136(2002)

Relationship between nucleotide sequence and 3D protein structure of six genes in *Escherichia coli*, by analysis of DNA sequence using a Markov model

Yuko Ohfuku [1, 2, 3*], Hideo Tanaka [2] and Masami Uebayasi [3]

[1] *National Institute of Evaluation and Technology 2-49-10 Nishihara Shibuyaku Tokyo, 151-0066, Japan*
[2] *Agricultural Sciences University of Tsukuba 1-1-1 Ten-nodai Tsukuba city Ibaraki, 305-8572, Japan*
[3] *Institute of Molecular and Cell Biology, AIST central 6, 1-1-1 Higashi Tsukuba Ibaraki, 305-8866, Japan*

**E-mail:ohfuku-yuko@meti.go.jp*

(Received June 3, 2002; accepted July 9, 2002; published online December 25, 2002)

Abstract

We speculate that structures of genes might have been produced not only by the origination of the gene by itself, but also by the combination of several parts of structures from other species, through events, such as gene transfer, fusion of DNA sequences, and rearrangement of DNA sequences, to create new functions in evolution. To study the relationship between the 3D structure and DNA sequence of proteins comprised of multi-domains, six genes in *Escherichia coli,* including those coding for Flavodoxin (ferredoxin) reductase, Flavin oxidoreductase, Integrase/recombinase *xer*D, Endonuclease III, Heat shock protein (*grp*E), and Elongation Factor Tu (*tuf*B), were chosen, since the 3D structures of these proteins had been determined by X-ray crystallography. The DNA sequences were analyzed by probability as calculated by using a species-specific inhomogeneous Markov model and divided into regions. The probability was calculated by the GeneMark program from the third-order matrix of Class III genes (*Escherichia coli* horizontally transferred genes) produced with the Markov model by Borodovsky M. et al.. The probability indicates the degree of similarity to the DNA sequence of the genes that were used for producing the matrix with the Markov model. Moreover, the probability indicates whether the origin of the divided DNA sequence would be different from that of the Class III genes. The divided regions, except for those of Flavin oxidoreductase, corresponded to the structural domains classified by CATH. Namely, the divided regions, suggested to be of different origins by analysis with the Markov model, showed correspondence to the domains and the gross content of the 3D structure. We can ascertain which DNA sequences would have been incorporated into the structure of a gene as parts of the structure in evolution. We propose a hypothesis that in evolution, the structure of genes might have been arisen by the incorporation of DNA sequences from other species.

Key Words: DNA sequence analysis, Markov model, evolution, structural evolution

Area of Interest: Bioinformatics and Bio Computing

1. Introduction

Some functional genes are composed of domains that are structurally separated into parts [1][2]. It is possible to consider that some of them might have been incorporated from other species as parts of the structure in the process of evolution. Therefore, we postulated that structures of genes might have been produced not only by the origination of the gene by itself, but as a composition of several parts of structures from other species, through such events as gene transfer, fusion of DNA sequences, and rearrangement of DNA sequences, etc., and in this way have created a new function in evolution. To demonstrate this hypothesis, we indicate how example DNA sequences of functional genes would be divided into structural domains. Scherer S. et al. reported that by analyzing patterns of probability of DNA sequences using the seventh-order Markov Models produced by DNA sequence fragments, the pattern of some region was found to be different from those of other regions [3]. They suggested the following three reasons to explain the difference of probability in the DNA sequence. One was that there might be a selection pressure on the chromosome. The second is that a lateral gene transfer might have occurred from another species. Last is that the DNA sequence might be incorporated from an organelle into the nuclear genome. They made no mention, however, about the relationship between the structure of the gene and the DNA sequence. Therefore, we thought that we could reveal the structure and function or evolution of a functional gene by analyzing its DNA sequence.

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, β-glucoside degradation, fucose degradation), genes containing insertion sequences, and genes behaving as mutators when inactivated (*mutH*, *mutT*, and *mutD*) and lambdoid phage lysogeny control protein. Also they suggested strongly that these Class III genes mostly comprised genes inherited by horizontal transfer [4][5]. 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. Accordingly, to explore the DNA sequence from other species, where structural information had been published, we used the third-order matrix, produced with Class III genes by Borodovsky M. et al. (*Escherichia coli* horizontally transferred genes) for our analysis [5]. 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. Moreover, the probability indicates that the origin of the divided DNA sequence would be different from that of the Class III genes.

To reveal the relationship between the 3D structure and DNA sequence for proteins comprised of multi-domains, six genes in *Escherichia coli,* such as Flavodoxin (ferredoxin) reductase, Flavin oxidoreductase, Integrase/recombinase *xer*D, Endonuclease III, Heat shock protein (*grp*E), and Elongation Factor Tu (*tuf*B), were chosen, because the 3D structures of these proteins had been determined by X-ray crystallography and the complete genomic DNA sequence of *Escherichia coli*

was determined. Since the probability calculated using a matrix of the genes produced with the Markov model indicates the degree of similarity to the DNA sequence of the genes, the DNA sequence of the genes could be analyzed by the probability. By comparison with the regions divided by the probability of the DNA sequence of the gene and the domain of the 3D structure, there was a correspondence between the divided regions of five genes excluding Flavin oxidoreductase in *Escherichia coli* and the domains classified by "Class", "Architecture", and "Topology" of CATH [6].

2. Materials and Methods

2.1 Materials

DNA sequences and 3D structures of Flavodoxin (ferredoxin) reductase [7], Flavin oxidoreductase [8], Integrase/recombinase *xer*D [9], Endonuclease III [10], Heat shock protein (*grp*E) [11], and Elongation Factor Tu (*tuf*B) [12] in *Escherichia coli* were downloaded from the Colibri Database [13] and Protein Data Bank [14], respectively.

2.2 Methods

The probability of a given segment of the DNA sequence is calculated by the program GeneMark. 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.) [15]. In our analysis, we 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. with 48 bases as the window size and three-base steps. The probabilities of DNA sequence were analyzed according the following definitions. Additionally, we compared our results with the domain classified by CATH.

2.3 Definition of class demonstrating the calculated probability

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 the 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 we grouped the probability scores into six classes by steps of 0.1, as given in Table 1.

Table1. 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

2.4 Definition of regions demonstrating classes

The probability was divided into regions when classes changed. But when the probability class ranges from class 3 to class 6, the frequency of the same probability class should be more than five. Also when the frequency of class 1 is more than four continuously, then that position or region was the boundary between two regions.

2.5 Definition of CLASS demonstrating classes of region

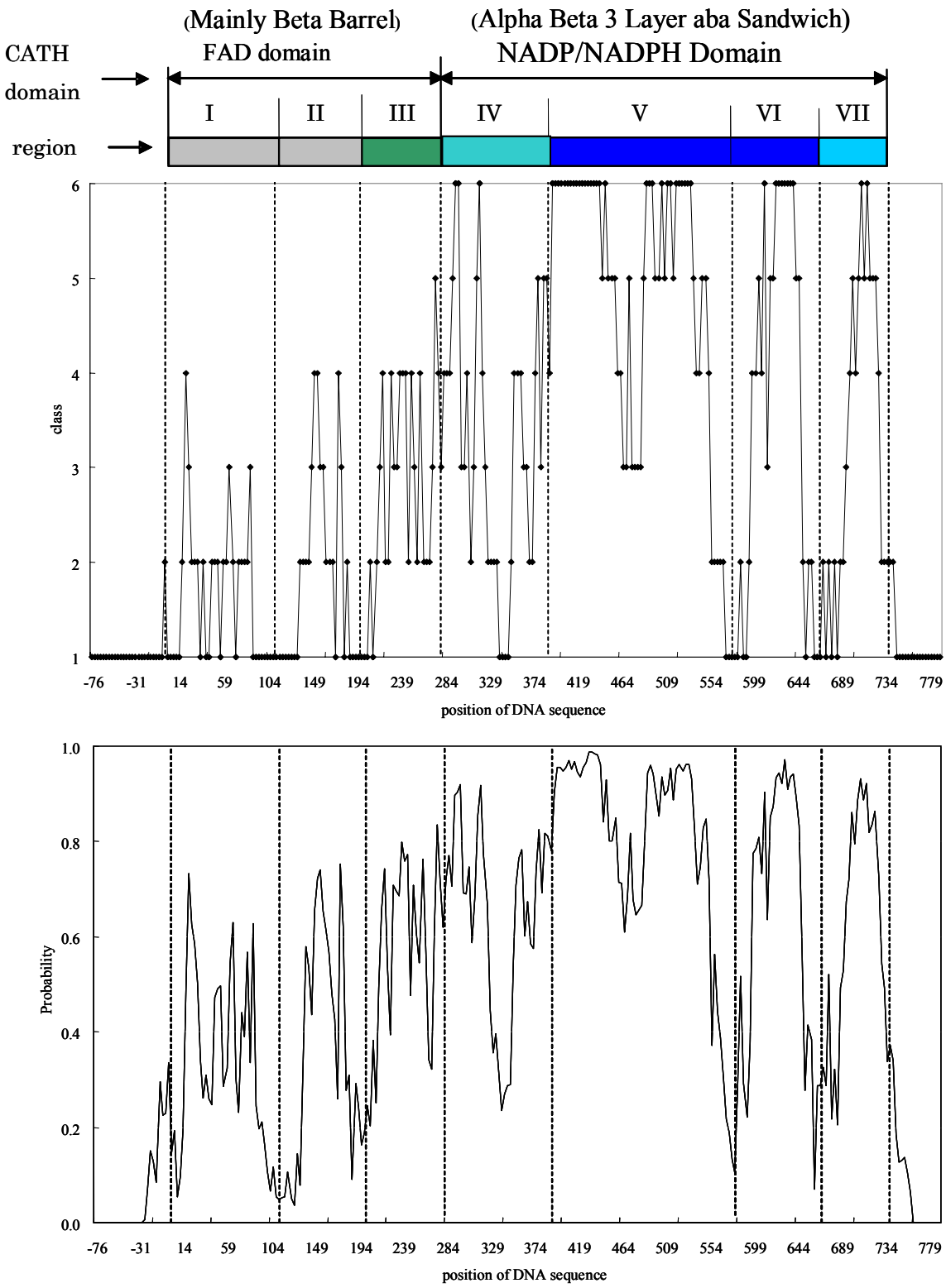
When the frequency of the same continuous class is more than five, the CLASS is regarded as a class, by itself. Conversely, when the frequency of the same continuous class is less than six, or if no class occurs continuously, then the most frequent class should be taken as the CLASS.

3. Results

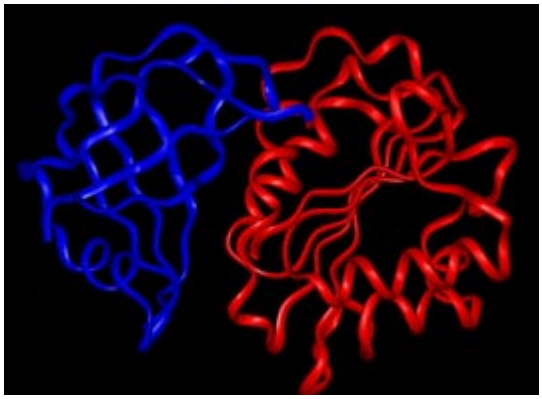
3.1 Analysis of probability of DNA sequence in Flavodoxin reductase

Fig. 1. (a) represents the plot of probabilities versus the position of the DNA sequence of Flavodoxin reductase and the plot between classes versus the position of the DNA sequence, classified by probability. The upper boxes indicate the divided regions and the domains classified by CATH. Fig. 1 (b) shows the 3D structure corresponding to the gene (PDB code 1fdr). The DNA sequence coding for Flavodoxin reductase could be divided into seven regions, from I to VII. Fig.1 (a) indicates the correspondence between the domains and the divided regions. Regions from I to III belong to the FAD domain, where Flavin Adenine Dinucleotide (FAD) binds [7]. The class and architecture of the FAD domain are “Mainly Beta” and “Barrel,” respectively, according to the definition of CATH. Regions from IV to VII belonged to the NADP domain, where NADP/NADPH binds [7]. The class and architecture of the NADP domain are “Alpha Beta” and “3 Layer (aba) Sandwich,” respectively. Each domain is classified based on “Class,” derived from the secondary structure, “Architecture,” which is derived from the gross orientation of the secondary structure, and “Topology” defined by CATH, and corresponding to the divided region.

(a) Regions divided by six classes of probability



(b) 3D structures of Flavodoxin reductase



(PDB code 1fdr)

Figure 1. Result of probability analysis of Flavodoxin reductase

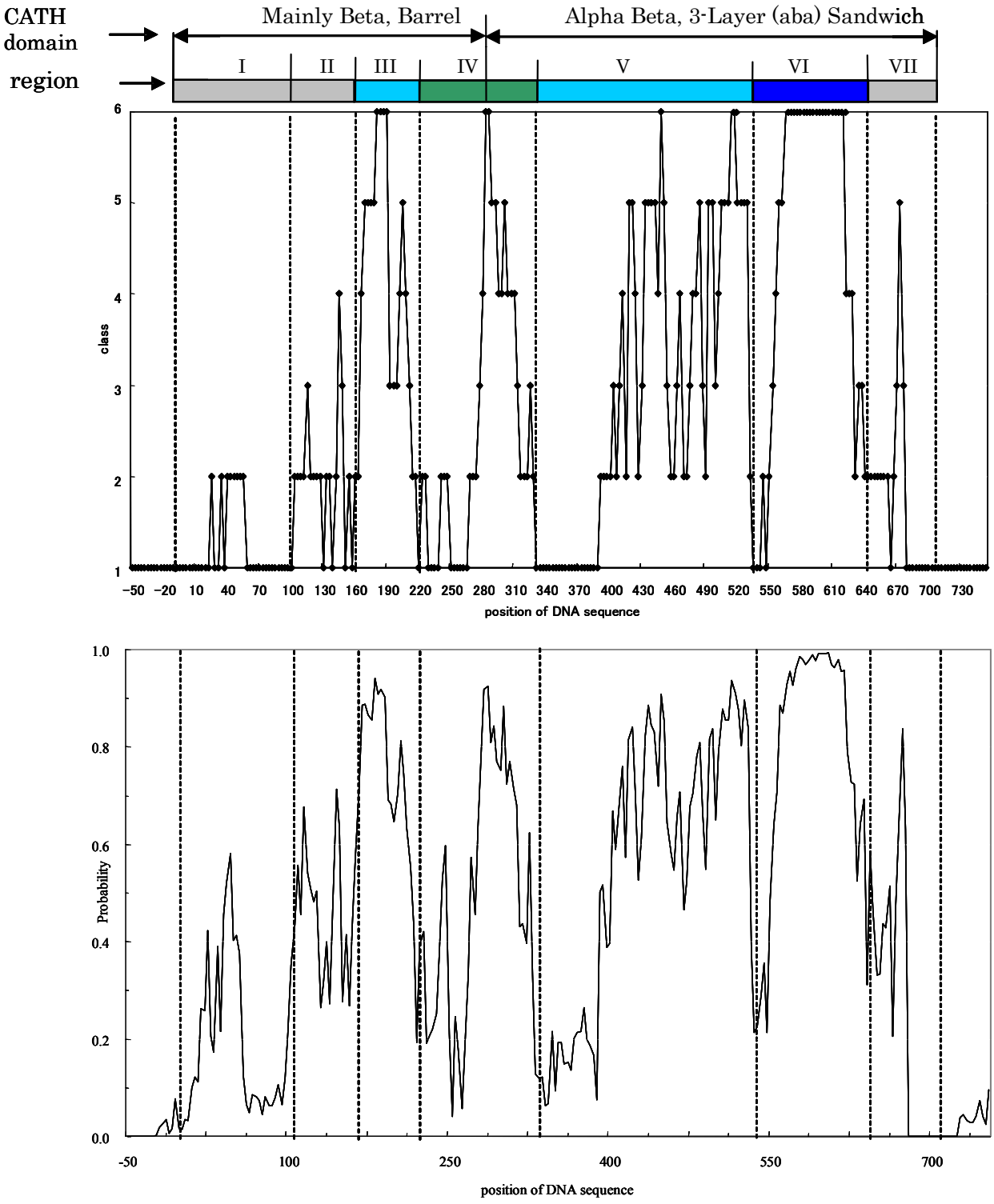
(a) shows the plot of probabilities versus the position of the DNA sequence of Flavodoxin reductase and the plot between classes versus the position of the DNA sequence, and classified by the probability calculated by the GeneMark program. The upper boxes indicate the divided regions and the domains classified by CATH. Light gray, dark green, cyan blue, and dark blue indicate CLASS 1+2, 4, 5, and 6, respectively.

(b) indicates the 3D structure of Flavodoxin reductase (PDB code 1fdr). Blue and red indicate the FAD domain and NADP/NADPH domain, respectively. Also blue and red indicate the two domains classified by CATH; one domain whose class is "Mainly Beta" and architecture is "Barrel," and the other domain whose class is "Alpha Beta" and architecture is "3-Layer (aba) Sandwich."

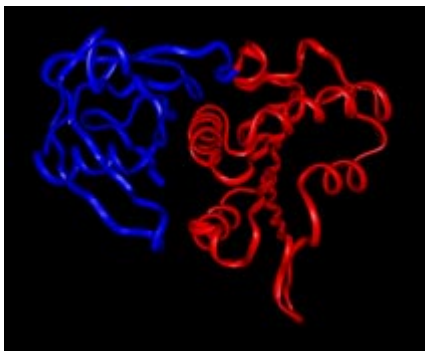
3.2 Analysis of probability of DNA sequence for Flavin oxidoreductase

The DNA sequence coding for Flavin oxidoreductase was analyzed by the same method as that used for Flavodoxin reductase. Fig. 2 (a) indicates the plots of probability and class, as well as region and domain of Flavin oxidoreductase similarly as shown in Fig. 1 (a). Fig 2 (b) shows the 3D structure corresponding to the gene (PDB code 1qfj). The DNA sequence coding for Flavin oxidoreductase could be divided into seven regions, from I to VII. Fig. 2 (a) indicates the correspondence with the domain and the divided region. The regions from I to the most part of IV belonged to the domain at the N-terminus, the class and architecture of which, in accordance with the definition of CATH, are "Mainly Beta" and "Barrel," respectively. The regions from a part of IV to VII belonged to the domain at the C-terminus, the class and architecture of which is "Mainly Alpha Beta" and "3-Layer (aba) Sandwich," respectively. In this scheme, however, region IV was located so that it bridged over both domains. Thus, this gene could not be completely divided into the regions corresponding to the structural domains.

(a) Regions divided by six classes of probability



(b) 3D structures of Flavin oxidoreductase



(PDB code 1qfj)

Figure 2. Result of analysis of probability of Flavin oxidoreductase

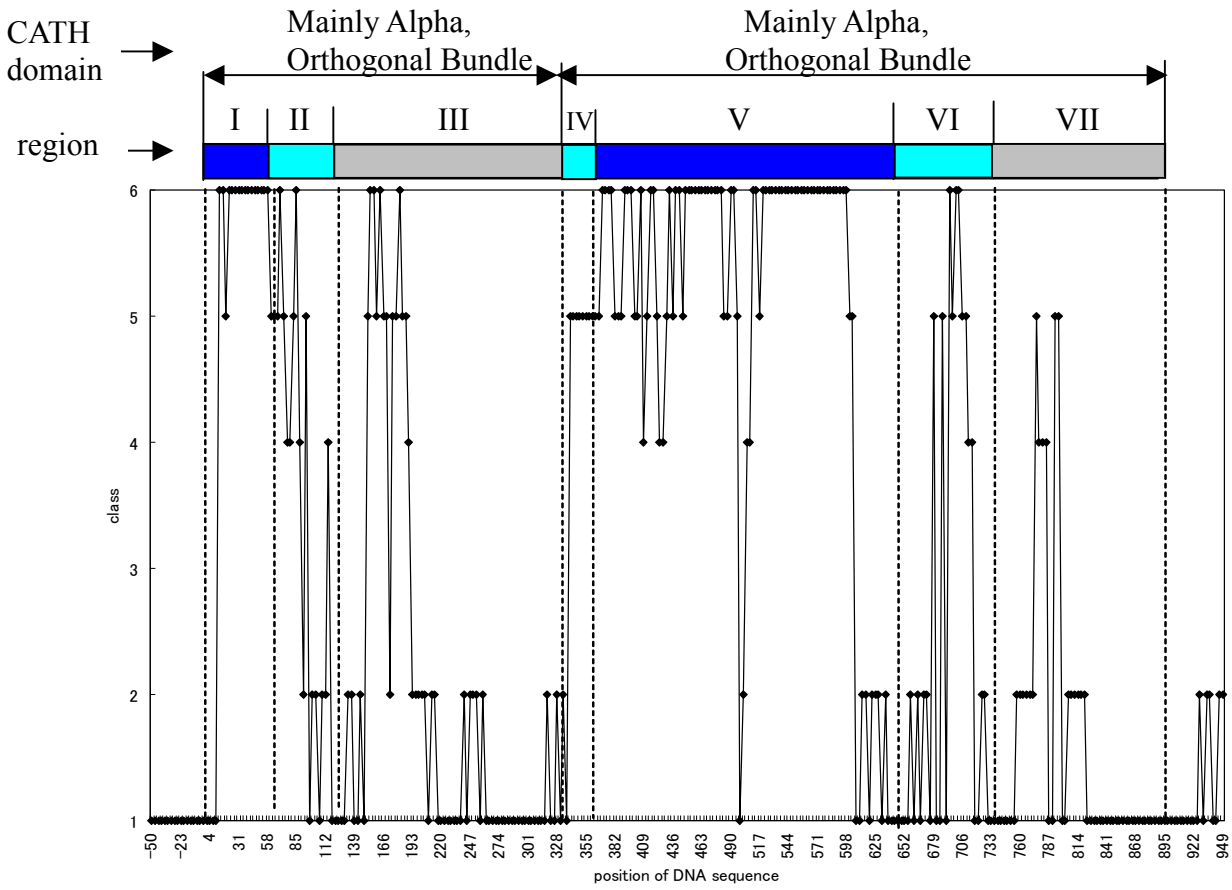
(a) shows the plots of probabilities versus the position of the DNA sequence for Flavin oxidoreductase and the plot between classes versus the position of the DNA sequence, as classified by the probability calculated by the GeneMark program. The upper boxes indicate the divided regions and the domains classified by CATH. Light gray, dark green, cyan blue, and dark blue indicate CLASS 1-2, 4, 5, and 6, respectively.

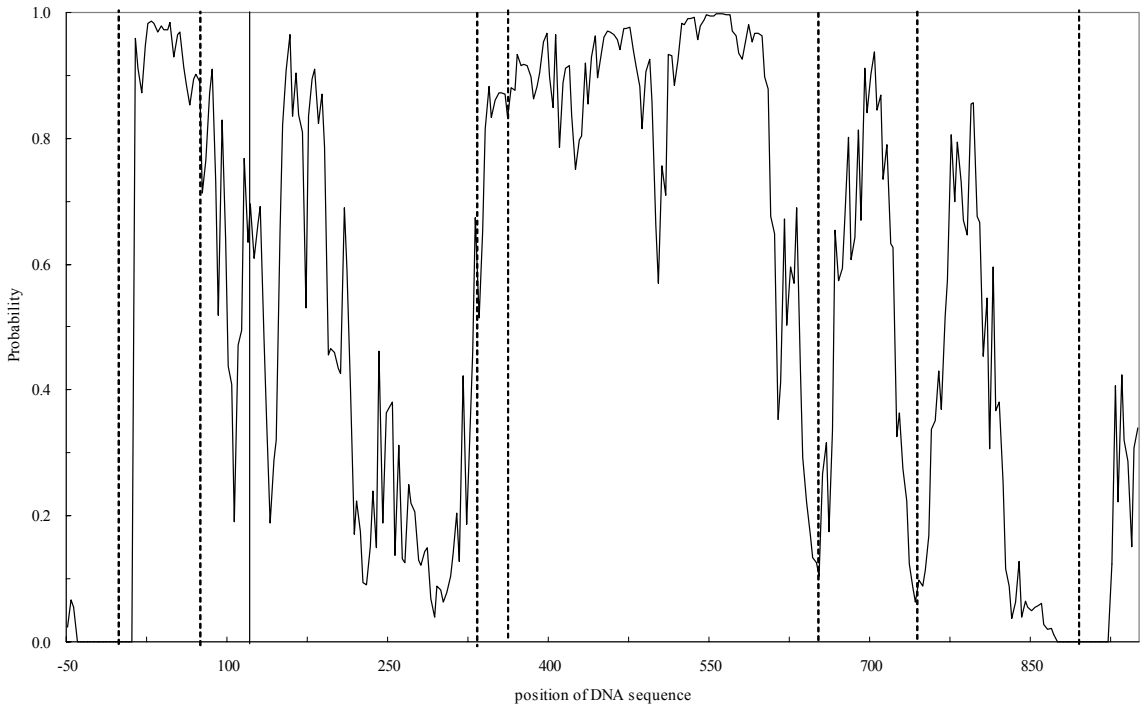
(b) indicates the 3D structure of Flavin oxidoreductase (PDB code 1qfj). Blue and red indicate the two domains classified by CATH; one domain whose class is "Mainly Beta" and architecture is "Barrel" and the other whose class is "Alpha Beta" and architecture is "3-Layer (aba) Sandwich."

3.3 Analysis of probability for the DNA sequence of Integrase/recombinase *xer*D

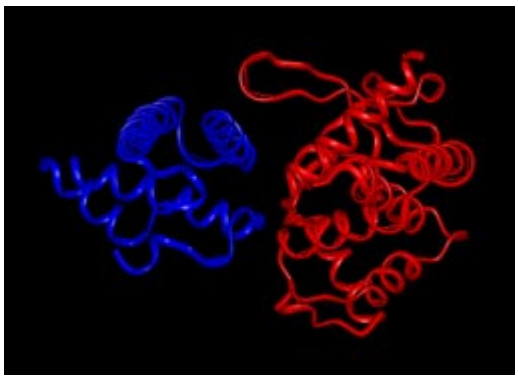
The DNA sequence coding for Integrase/recombinase *xer*D was analyzed by the same method as that used for Flavodoxin reductase. Fig. 3 (a) indicates the plots of probability and class, and region and domain of Integrase/recombinase *xer*D similarly as those of Fig. 1 (a). Fig 3 (b) shows the 3D structure corresponding to the gene (PDB code 1a0p). The DNA sequence for Integrase/recombinase *xer*D could be divided into seven regions, from I to VII. Fig. 3 (a) indicates the correspondence with the domain and the divided region. The regions from I to III belonged to the domain at the N-terminus, the class and architecture of which, in accordance with the definition of CATH, are "Mainly Alpha" and "Orthogonal Bundle," respectively. The regions from IV to VII belonged to the domain at the C-terminus, the class and architecture of which, are "Mainly Alpha" and "Orthogonal Bundle," respectively. Each domain classified based on the "Class," derived from the secondary structure, "Architecture," which is derived from the gross orientation of the secondary structure, and "Topology" defined by CATH, corresponded with the divided region.

(a) Regions divided by six classes of probability





(b) 3D structures of Integrase/recombinase *xer*D



(PDB code 1aOp)

Figure 3. Result of analysis of probability of Integrase/recombinase *xer*D

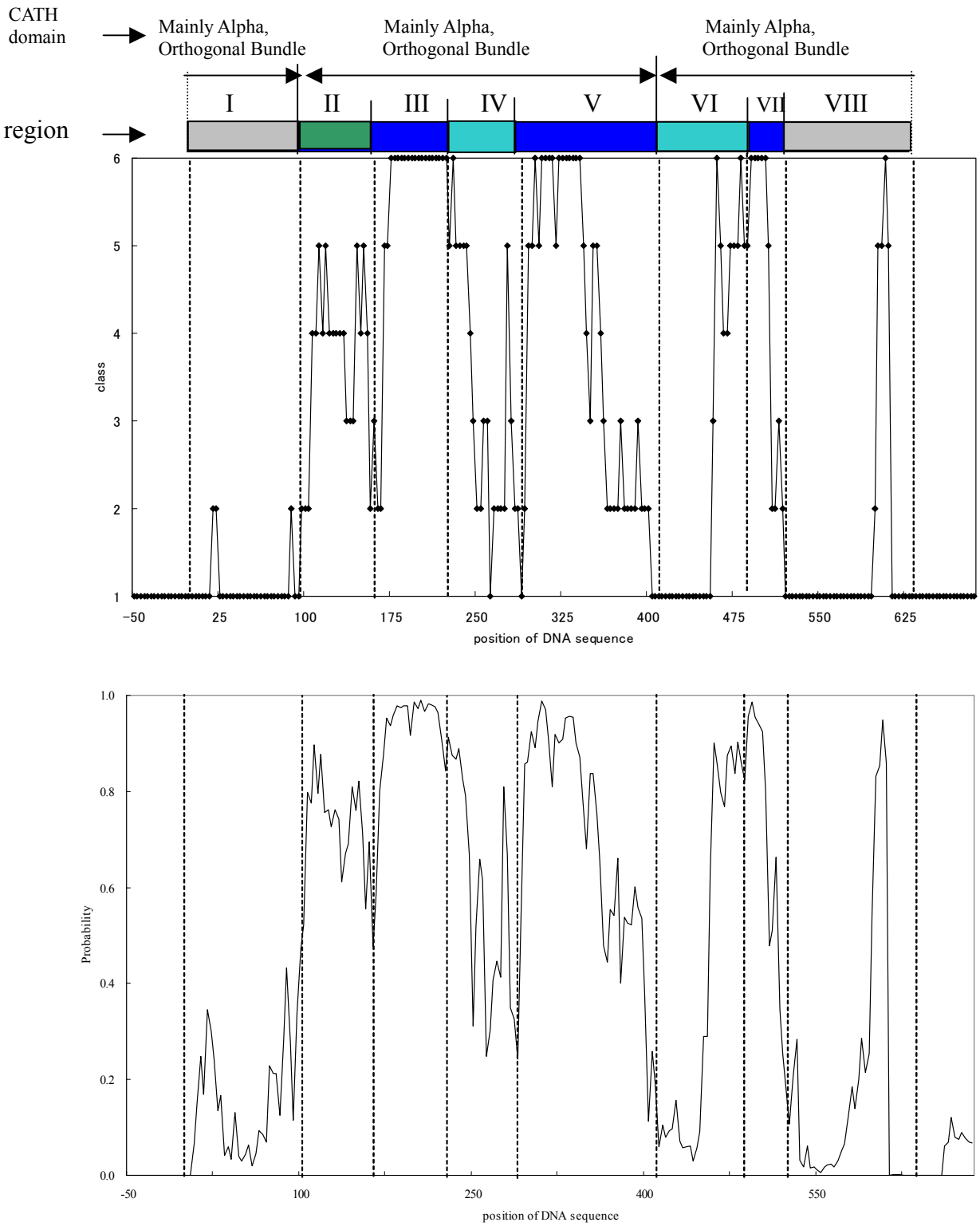
(a) shows the plot of probabilities versus the position of the DNA sequence for Integrase/recombinase *xer*D and the plot between classes versus the position of the DNA sequence, classified by probability as calculated by the GeneMark program. The upper boxes indicate the divided regions and the domains classified by CATH. Light gray, dark green, cyan blue, and dark blue indicate CLASS 1+2, 4, 5, and 6, respectively.

(b) indicates the 3D structure of Integrase/recombinase *xer*D (PDB code 1a0p). Blue and red indicate the two domains classified by CATH; the class and architecture of both domains are "Mainly Alpha" and "Orthogonal Bundle."

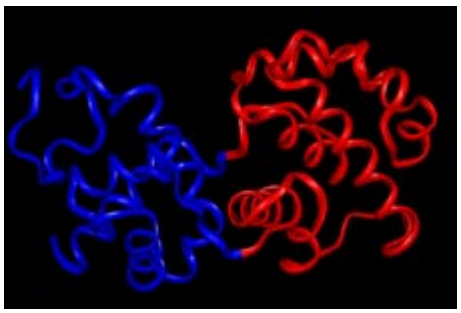
3.4 Analysis of probability of DNA sequence in Endonuclease III

The DNA sequence coding for Endonuclease III was analyzed by the same method as that for Flavodoxin reductase. Fig. 4 (a) indicates the plots of probability and class, and region and domain of Endonuclease III as well as those of Fig. 1 (a). Fig 4 (b) shows the 3D structure corresponding to the gene (PDB code 2abk). The DNA sequence of Endonuclease III could be divided into eight regions, from I to VIII. Fig.4 (a) indicates the correspondence with the domain and the divided region. Regions I and from VI to VIII belonged to the domain at the N-terminus, class and architecture of which, according to the definition of CATH, are "Mainly Alpha" and "Orthogonal Bundle," respectively. Regions from II to V belonged to the domain at the C-terminus, whose class and architecture are "Mainly Alpha" and "Orthogonal Bundle," respectively. Each domain was classified based on "Class," derived from the secondary structure, "Architecture," which was derived from the gross orientation of the secondary structure and "Topology" as defined by CATH, corresponded with the divided region.

(a) Regions divided by six classes of probability



(b) 3D structures of Endonuclease III



(PDB code 2abk)

Figure 4. Result of analysis of probability of Endonuclease III

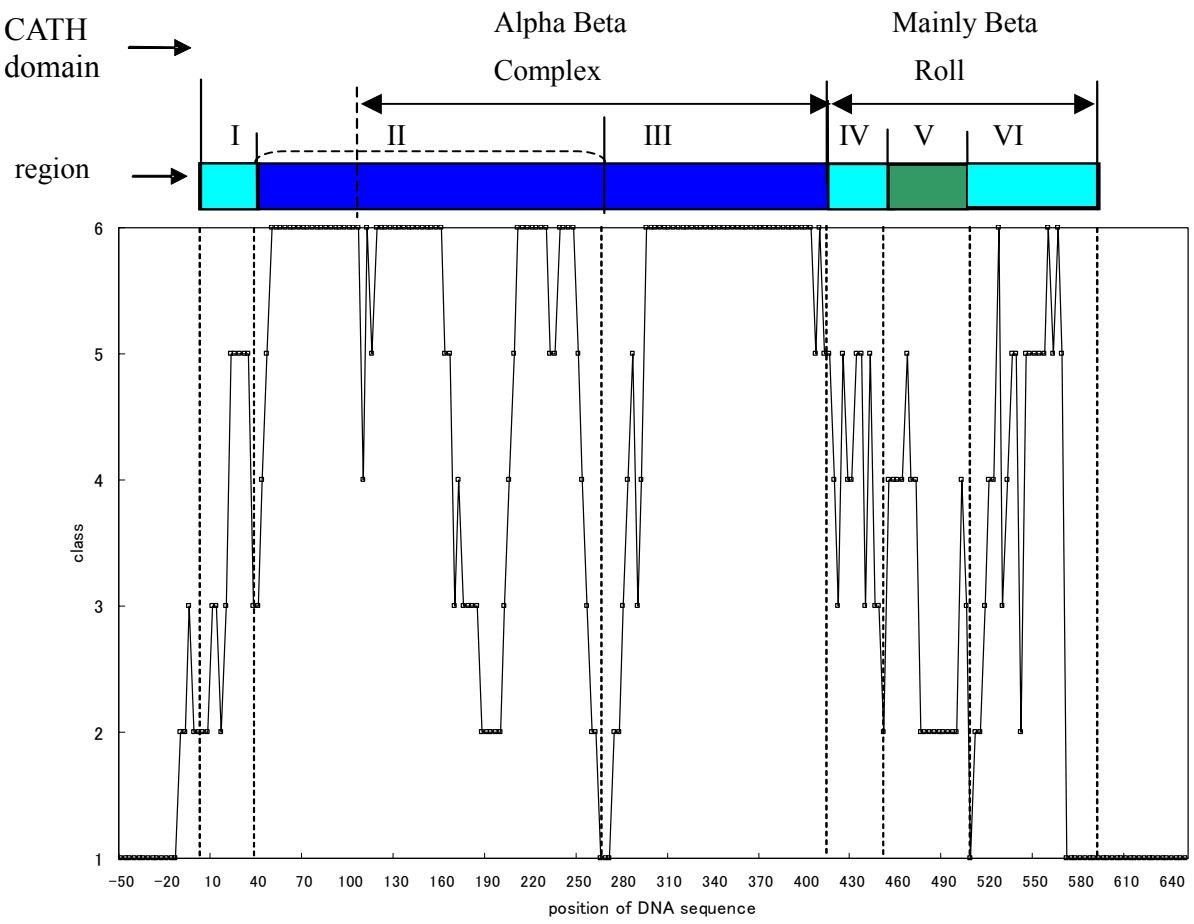
(a) shows the plot of probabilities versus the position of the DNA sequence for Endonuclease III and the plot between classes versus the position of the DNA sequence, as classified by the probability calculated by the GeneMark program. The upper boxes indicate the divided regions and the domains classified by CATH. Light gray, dark green, cyan blue, and dark blue indicate CLASS 1-2, 4, 5, and 6, respectively.

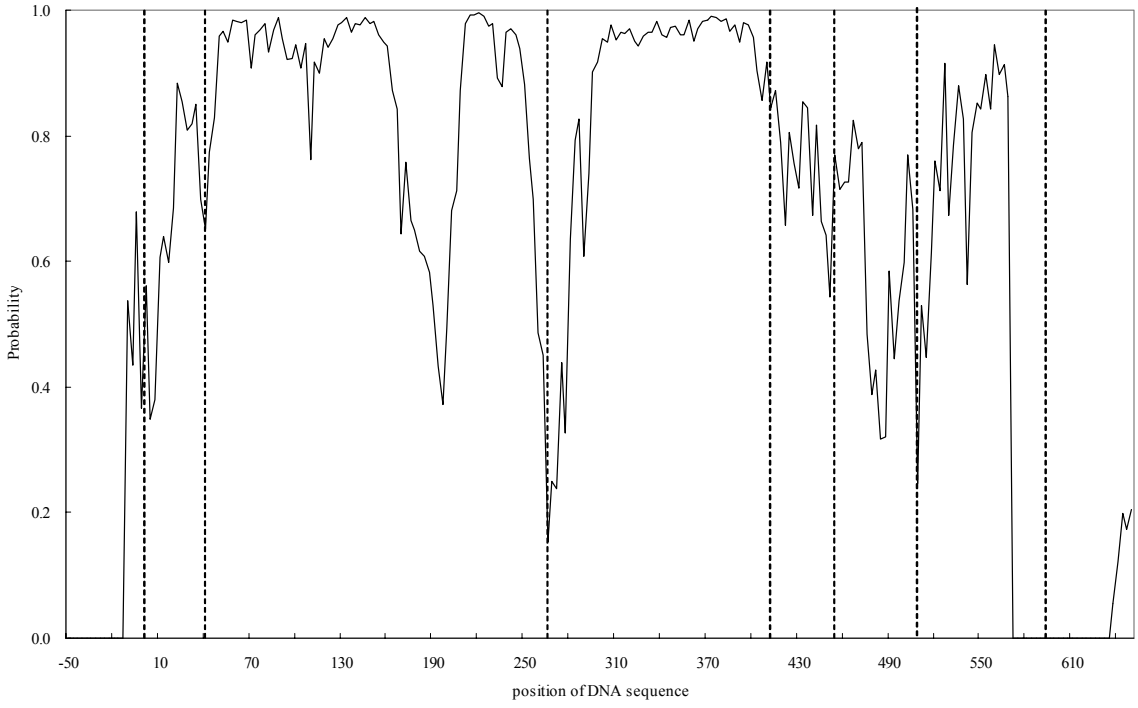
(b) indicates the 3D structure of Endonuclease III (PDB code 2abk). Blue and red indicate the two domains classified by CATH, the class and architecture of both domains are "Mainly Alpha" and "Orthogonal Bundle."

3.5 Analysis of probability of the DNA sequence in Heat shock protein (*grp*E)

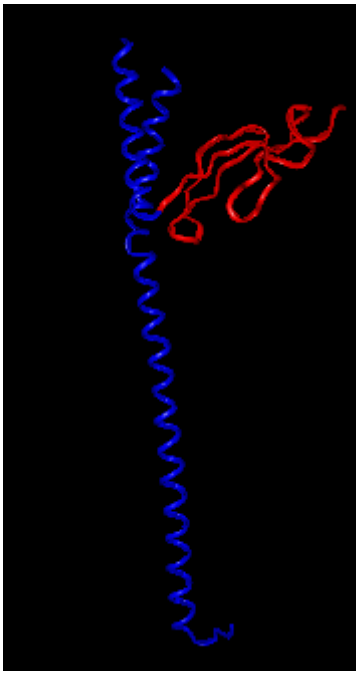
The DNA sequence of Heat shock protein (*grp*E) was analyzed by the same method as that for Flavodoxin reductase. Fig. 5 (a) indicates the plots of probability and class, and region and domain of Heat shock protein (*grp*E) similarly as those of Fig. 1 (a). Fig. 5 (b) shows the 3D structure coded for by the gene (PDB code 1dkg). The DNA sequence of Heat shock protein (*grp*E) could be divided into six regions, from I to VI. Fig. 5 (a) indicates the correspondence with the domain and the divided region. Regions I, II, and III belonged to the domain at the N-terminus, 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. Each domain was classified based on "Class," derived from the secondary structure, "Architecture," which was derived from the gross orientation of the secondary structure, and "Topology" as defined by CATH corresponded with the divided region.

(a) Regions divided by six classes of probability





(b) 3D structures of Heat shock protein (*grp*E)



(PDB code 1dkg)

Figure 5. Result of the analysis of probability of Heat shock protein (*grp*E)

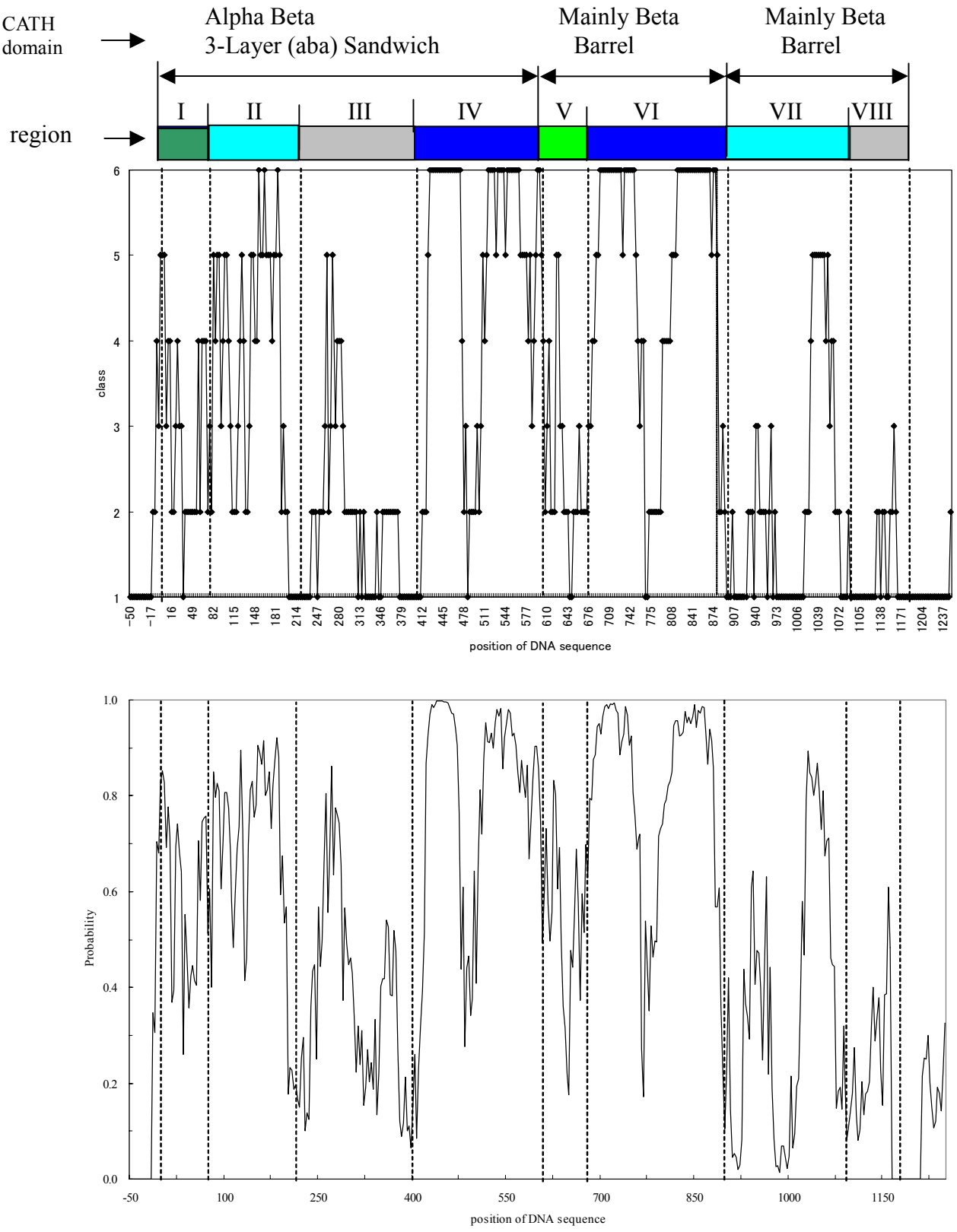
(a) shows the plot of probabilities versus the position of the DNA sequence for Heat shock protein (*grp*E) and the plot between classes versus the position of the DNA sequence, as classified by the probability calculated by the GeneMark program. The upper boxes indicate the divided regions and the domains classified by CATH. Dark green, cyan blue, and dark blue indicate CLASS 4, 5, and 6, respectively.

(b) indicates the 3D structure of Heat shock protein (*grp*E) (PDB code 1dkg). Blue and red indicate the two domains classified by CATH. The class of one domain is "Alpha Beta" and the architecture is "Complex," the class of the other domain is "Mainly Beta" and the architecture is "Roll."

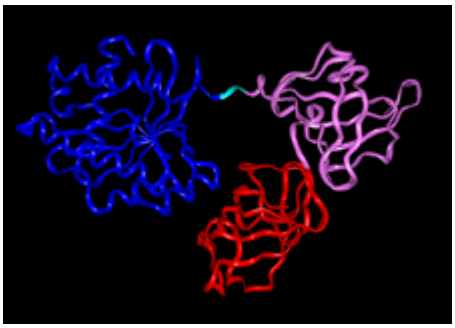
3.6 Analysis of probability of the DNA sequence in Elongation Factor Tu (*tuf*B)

The DNA sequence of Elongation Factor Tu (*tuf*B) was analyzed by the same method as that used for Flavodoxin reductase. Fig. 6 (a) indicates the plots of probability and class, and region and domain of Elongation Factor Tu similarly as those of Fig. 1 (a). Fig. 6 (b) shows the 3D structure coded for by the gene (PDB code 1efu). The DNA sequence of Elongation Factor Tu (*tuf*B) could be divided into eight regions, from I to VIII. Fig. 6 (a) indicates the correspondence with the domain and the divided region. The regions from I to IV belong to the domain at the N-terminus, the class and architecture of which, in accordance with the definition of CATH are "Alpha Beta" and "3-Layer (aba) Sandwich," respectively. Regions V and VI belong to the domain between the N-terminus and the C-terminus, the class and architecture of which are "Mainly Beta" and "Barrel," respectively. Regions VII and VIII belong to the domain at the C-terminus, and have the class and architecture of "Mainly Beta" and "Barrel," respectively. Each domain, which is classified on the basis of "Class" as derived from the secondary structure, "Architecture," which is derived from the gross orientation of the secondary structure, and "Topology" defined by CATH, corresponded with the divided region.

(a) Regions divided by six classes of probability



(b) 3D structures of Elongation Factor Tu



(pdb code 1efu)

Figure 6. Result of analysis of the probability of Elongation Factor Tu (*tuf*B)

(a) shows the plot of probabilities versus the position of the DNA sequence of Elongation Factor Tu (*tuf*B) and the plot between classes versus the position of the DNA sequence, classified by the probability calculated by the GeneMark program. The upper boxes indicate the divided regions and the domains classified by CATH. Light gray, greenish yellow, dark green, cyan blue, and dark blue indicate CLASS 1+2, 3, 4, 5, and 6, respectively.

(b) shows the 3D structure of Elongation Factor Tu (PDB code 1efu). Dark blue, pink, and red indicate the three domains classified by CATH, the class of the domain at the N-terminus is "Alpha Beta" and the architecture is "3-Layer (aba) Sandwich." For the remaining two domains, the class is "Mainly Beta" and the architecture is "Barrel."

3.7 Comparison between the divided regions and the domain classified in accordance with the definition of CATH

Table 2 indicates the summary of the comparison between the divided regions and the domain classified in accordance with the definition of CATH. The divided regions of five genes, except for Flavin oxidoreductase in *Escherichia coli* by the probability calculated using Markov models, corresponded 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.

Table 2. Comparison between the divided regions and the domains classified in accordance with the definitions of CATH

Gene (PDB code9	Number of domains	CATH code of domains (class, architecture)	Number of regions
Flavodoxin reductase (1fdr)	2	Mainly beta, Barrel (FAD domain)	I-III
		Alpha beta, 3-Layer(aba) Sandwich (NADP/NADPH domain)	IV-VII
Flavin oxidoreductase (1qfj)	2	Mainly beta, Barrel	I-IV
		Alpha beta, 3-Layer(aba) Sandwich	IV-VII
Integrase/recombinase *xer*D (1a0p)	2	Mainly alpha, Orthogonal Bundle	I-III
		Mainly alpha, Orthogonal Bundle	IV-VII
Endonuclease III (2abk)	2	Mainly alpha, Orthogonal Bundle	I, VI-VIII
		Mainly alpha, Orthogonal Bundle	II-V
Heat shock protein (*grp*E)(1dkg)	2	Alpha beta, Complex	II, III
		Mainly beta, Roll	IV-VI
Elongation Factor Tu (*tuf*B) (1efu)	3	Alpha beta, 3-Layer (aba) Sandwich	I-IV
		Mainly beta, Barrel	V, VI
		Mainly beta, Barrel	VII, VIII

4. Discussion

Six 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), were analyzed. Five of the six 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 VII 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. By our analysis, however, the domain classified by CATH could be further divided into several more regions. But only region VI of Flavin oxidoreductase was bridged between two domains. Looking into more details, we could include one helix into the former domain among the two domains classified by CATH. We also noted that certain genes exist, such as that for Flavin oxidoreductase, which 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, we were 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.

Although we will demonstrate the results of our analysis more clearly in next paper, we suggest that the DNA sequence of the gene could be divided into the regions, where gene transfer and recombination would have occurred. This is because the divided regions, which showed correspondence to the domains and the gross 3D protein structure, would be elucidated to have different origins by analysis of the probability with the Markov model. Therefore, it is clear that these five genes would be composed of the DNA sequences, which might have been collected from parts of domains of other species or by recombination within its genome. We concluded that these five genes would have been produced by means of such events as horizontal transfer, rearrangement, or recombination of its own DNA sequences or those from other species in the process of evolution.

References

[1] Y. Liu and D. Eisenberg, *Protein*. *Sci*., **11**, 1285-1299 (2002).
[2] M. Jaskolski, *Acta*. *Biochim*. *Pol*., **48**, 807-827 (2001).
[3] S. Scherer et al., *Proc. Natl*. *Acad*. *Sci*., **91**, 7134-7138 (1994).
[4] C. Médigue et al., *J. Mol*. *Biol*., **222**, 851-856 (1991).
[5] M. Borodovsky et al., *Nucleic Acids*. *Res*., **23**, 3554-3562 (1995).
[6] C. A. Orengo et al., *Structure*, **5**, 1093-1108 (1997).
[7] M. Ingelman et al., *J. Mol*. *Biol*., **268**, 147-157 (1997).
[8] M. Ingelman et al., *Biochemistry*, **38**, 7040-7049(1999).
[9] H. S. Subramanya et al., *EMBO J*., **16**, 5178-5187 (1997).
[10] C. F. Kuo et al., *Science*, **258**, 434-440 (1992).
[11] C. J. Harrison et al., *Science*, **276**, 431-435 (1997).
[12] T. Kawashima et al., *Nature*, **379**, 511-518 (1996).
[13] http://genolist.pasteur.fr/Colibri/
[14] http://www.rcsb.org/pdb/
[15] http://opal.biology.gatech.edu/GeneMark/