Chapter 4

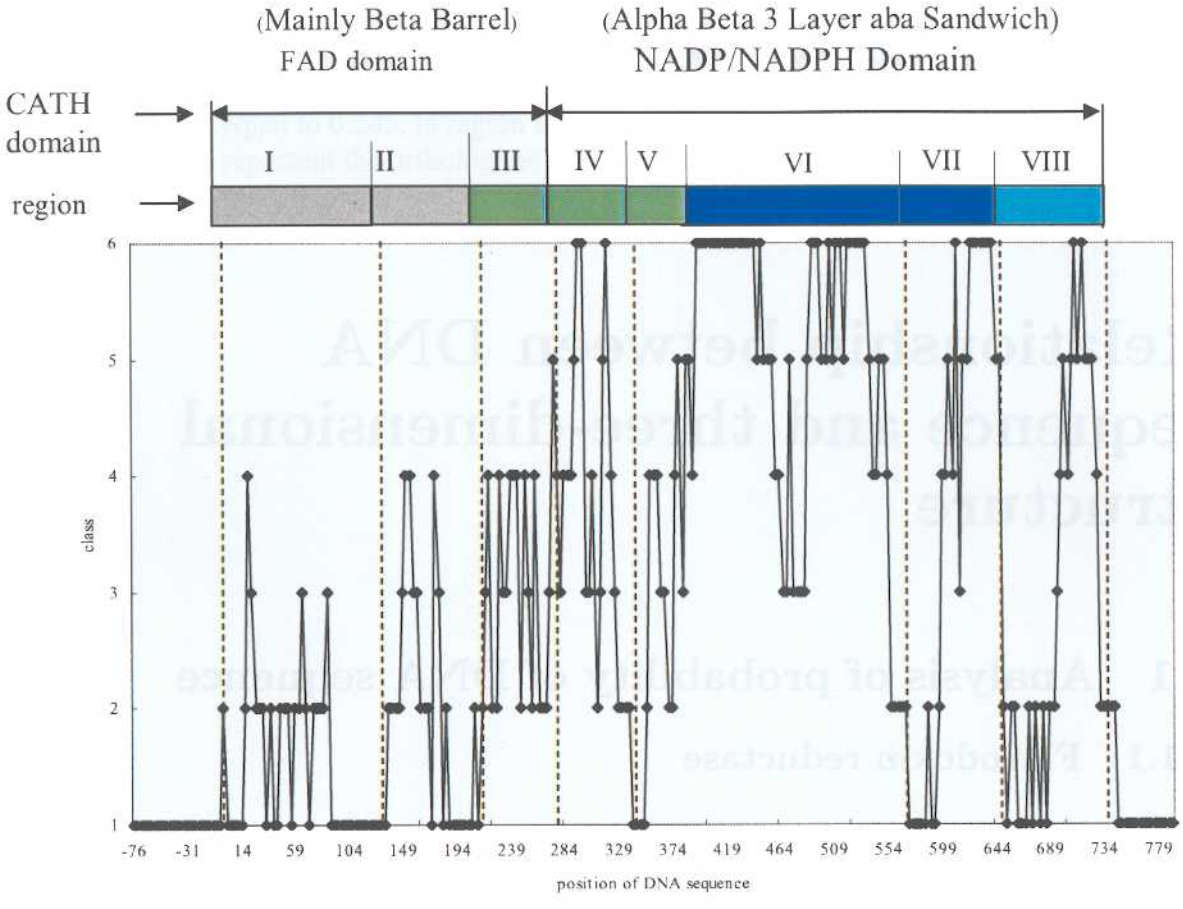
Relationship between DNA sequence and three-dimensional structure

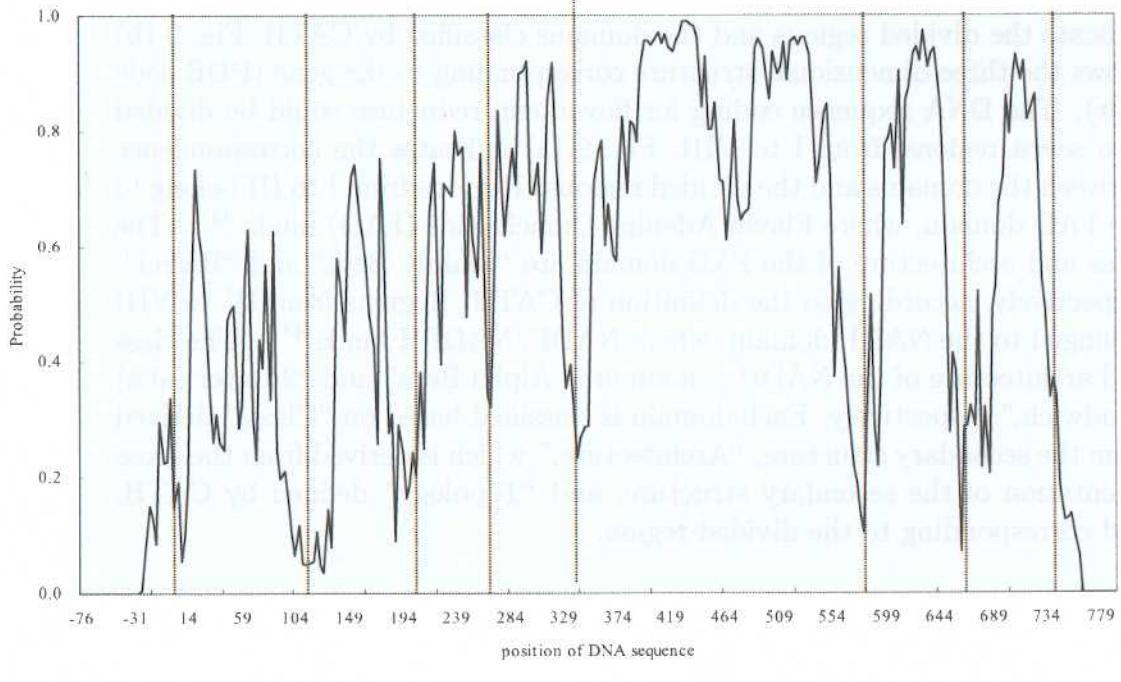
4.1 Analysis of probability of DNA sequence

4.1.1 Flavodoxin reductase

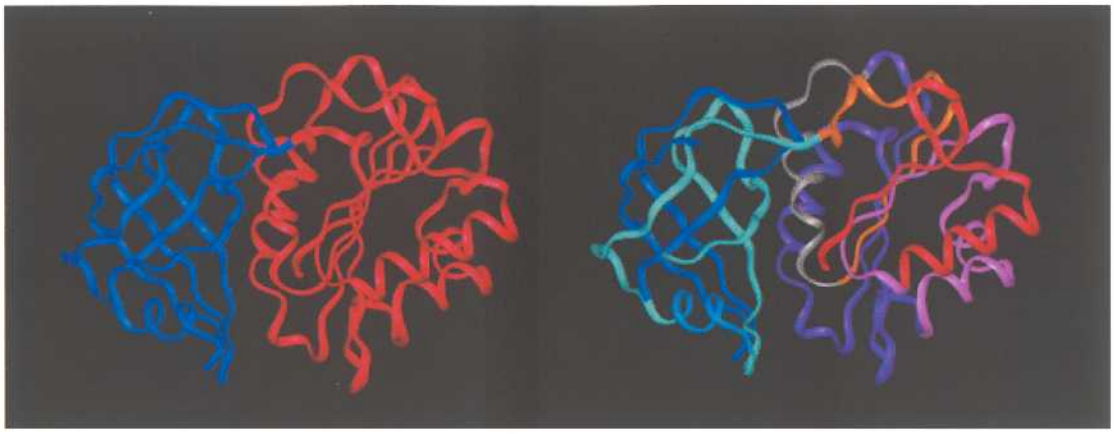
Fig. 9 (a) represents the plot of probabilities versus the position of the DNA sequence of flavodoxin reductase and the plot between classes versus the position of the DNA sequence, classified by probability. The upper boxes indicate the divided regions and the domains classified by CATH. Fig. 9 (b) shows the three-dimensional structure corresponding to the gene (PDB code 1fdr). The DNA sequence coding for flavodoxin reductase could be divided into seven regions, from I to VIII. Fig. 9 (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 [43] . The class and architecture of the FAD domain are "Mainly Beta" and "Barrel," respectively, according to the definition of CATH. Regions from IV to VIII belonged to the NADP domain, where NADP/NADPH binds [43] . 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) The three-dimensional structure of flavodoxin reductase



(PDB code: 1fdr)

Figure 9 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 the CLASS I-II, IV, V, and VI, respectively.

(b) indicates the three-dimensional structure of flavodoxin reductase (PDB code 1fdr). The figure on the left indicates the domains classified by CATH. 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". The figure on the right shows the seven regions divided by probability. Dark blue, light blue, sky blue, orange, gray, violet, pink and red represent regions, I, II, III, IV, V, VI, VII, and VIII, respectively.

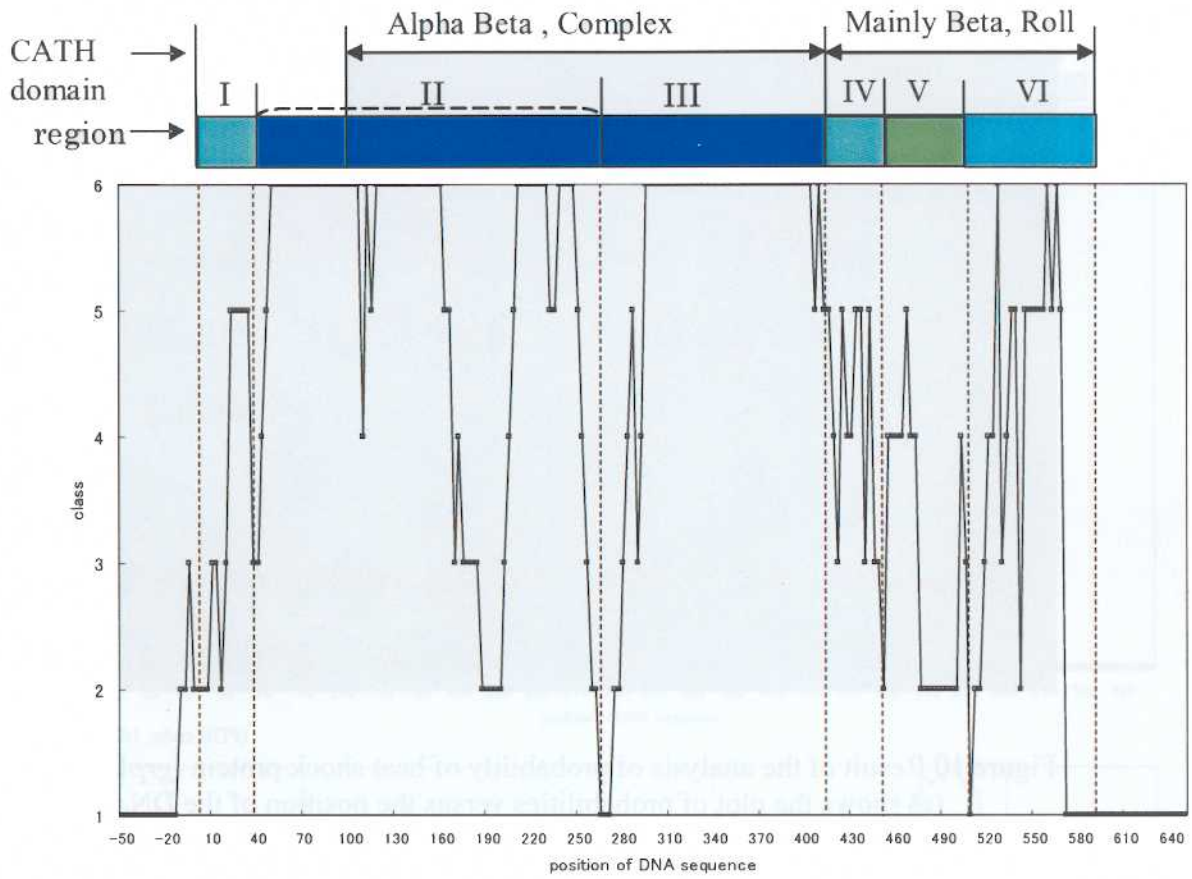
4.1.2 Heat shock protein (*grp*E)

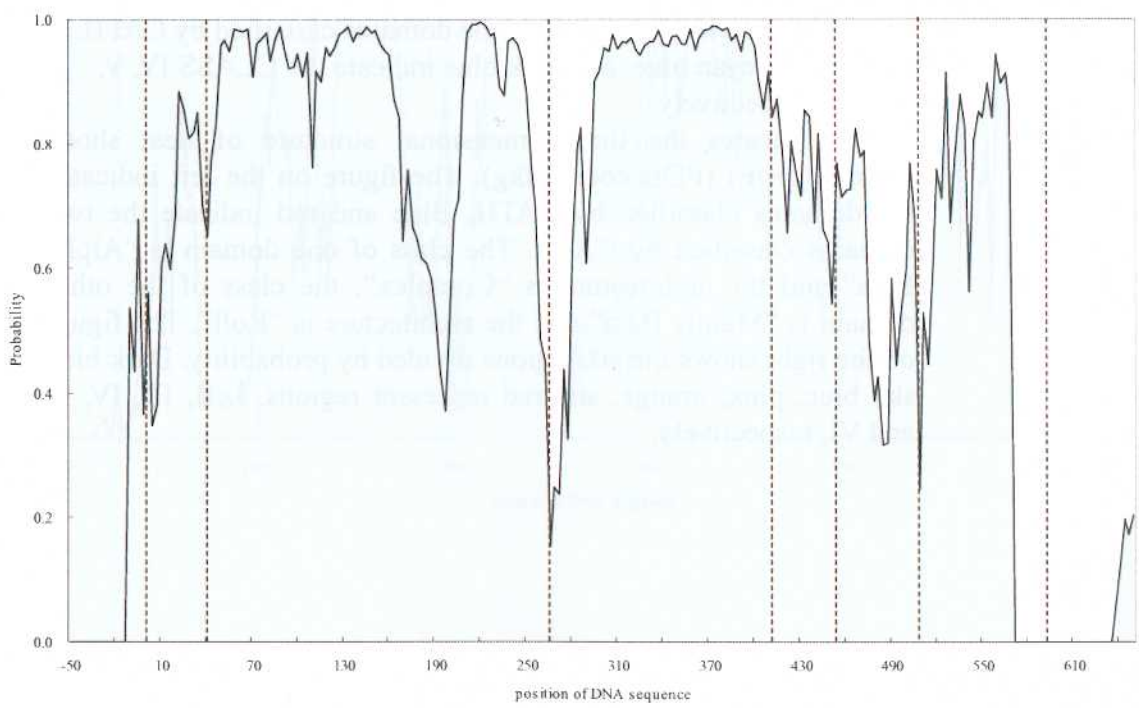
The DNA sequence of heat shock protein (*grp*E) was analyzed by the same method as that for flavodoxin reductase. Fig. 10 (a) indicates the plots of probability and class, and region and domain of heat shock protein (*grp*E) similarly as those of Fig. 9 (a). Fig. 10 (b) shows the three-dimensional structure 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. 10 (a) indicates the correspondence with the domain and the divided region. Approximately three-quarters from the end of region II and 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. 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.

4.1.3 Flavin oxidoreductase

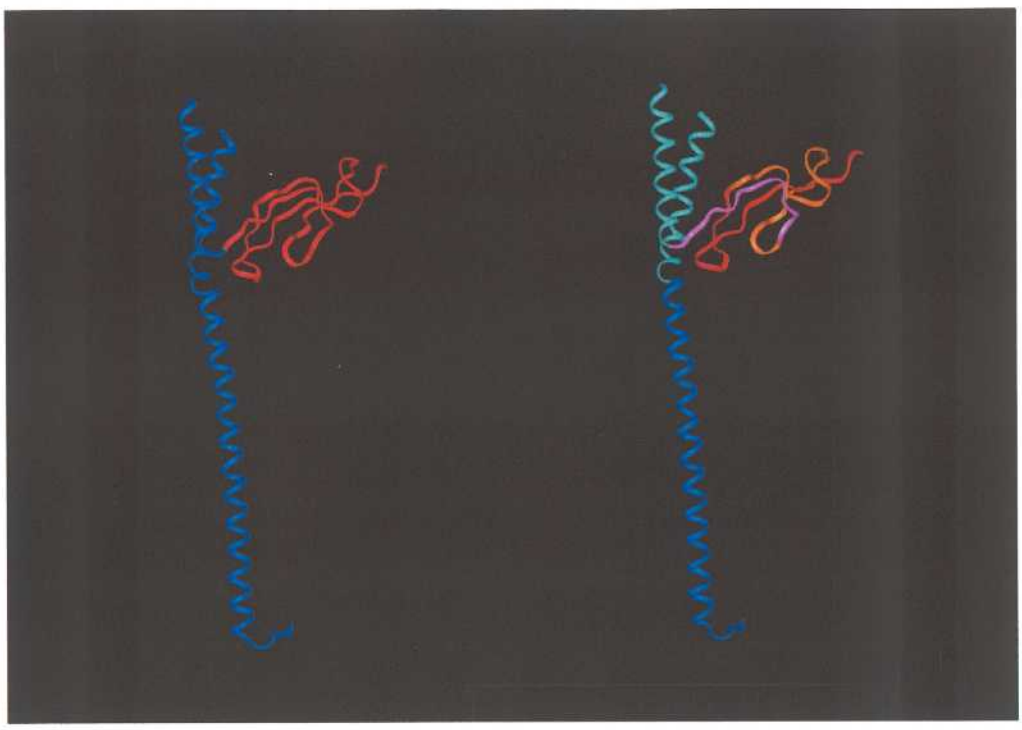
The DNA sequence coding for flavin oxidoreductase was analyzed by the same method as that used for flavodoxin reductase. Fig. 11 (a) indicates the plots of probability and class, as well as region and domain of flavin oxidoreductase similarly as shown in Fig. 9 (a). Fig 11 (b) shows the three-dimensional 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. 11 (a) indicates the correspondence with the domain and the divided region. The regions from I to the most part of III 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 III 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 III 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) The three-dimensional structure of Heat shock protein (*grp*E)



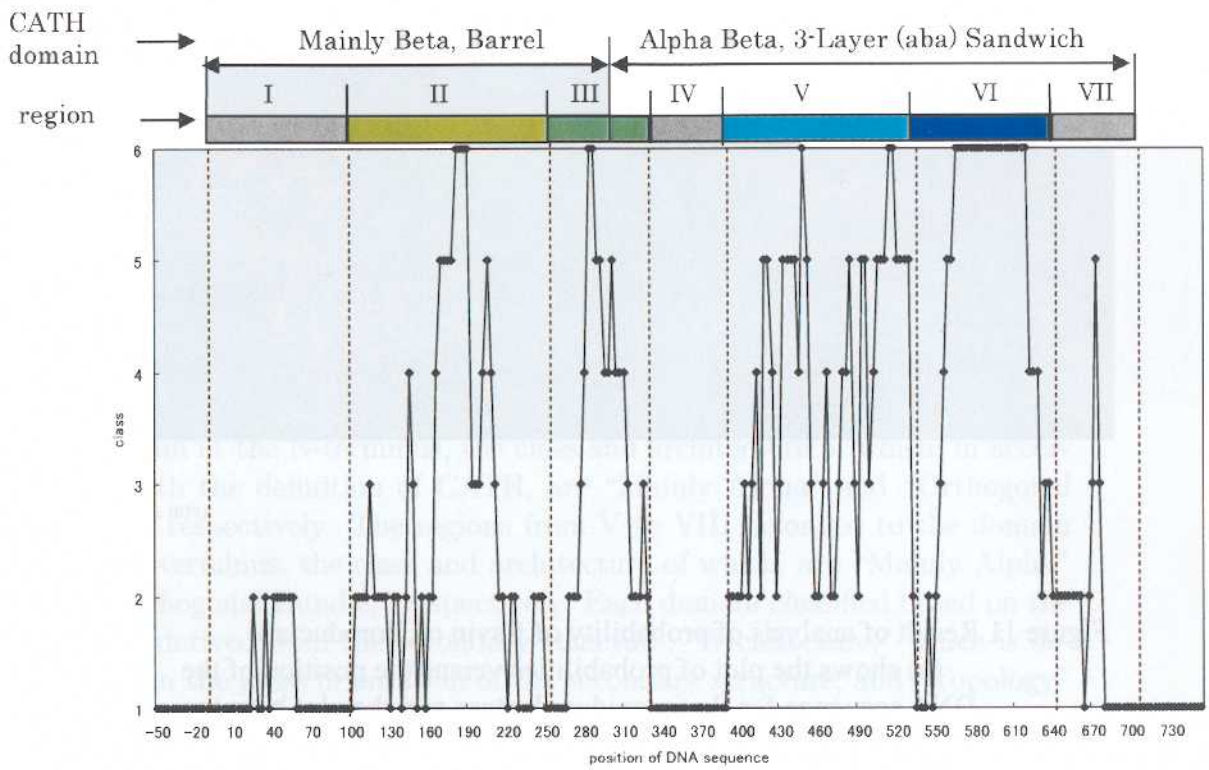
(PDB code: 1dkg)

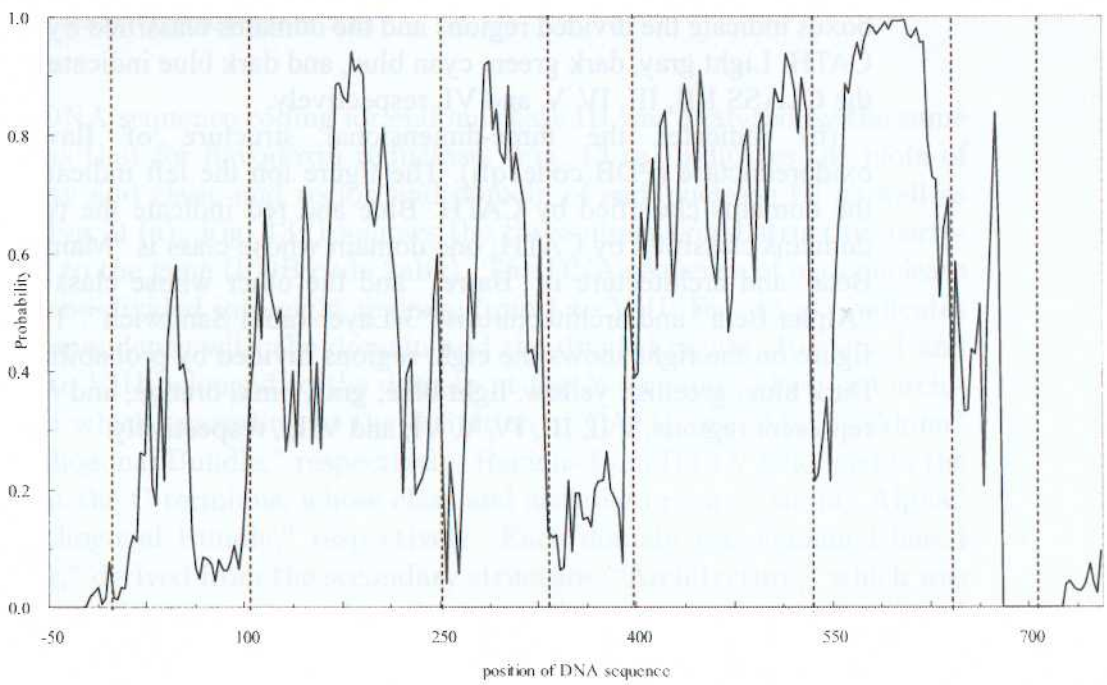
Figure 10 Result of the analysis of probability of heat shock protein (*grp*E)

(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 the CLASS IV, V, and VI, respectively.

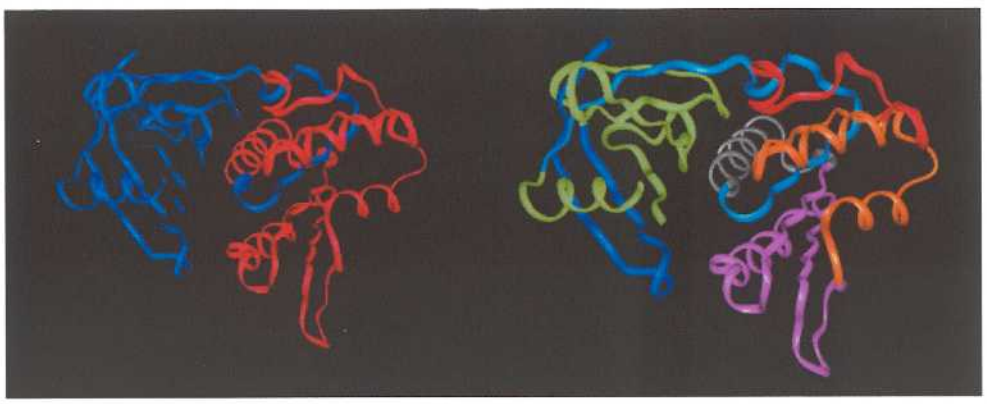
(b) indicates the three-dimensional structure of heat shock protein (*grp*E) (PDB code 1dkg). The figure on the left indicates the domains classified by CATH. Blue and red indicate the two domains classified by CATH. The class of one domain is "Alpha Beta" and the architecture is "Complex"; the class of the other domain is "Mainly Beta" and the architecture is "Roll". The figure on the right shows the six regions divided by probability. Dark blue, sky blue, pink, orange, and red represent regions, I, II, III, IV, V, and VI, respectively.

(a) Regions divided by six classes of probability





(b) The three-dimensional structure of flavin oxidoreductase



(PDB code: 1qfj)

Figure 11 Result of analysis of probability of flavin oxidoreductase

(a) shows the plot 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 the CLASS I-II, III, IV, V, and VI, respectively.

(b) indicates the three-dimensional structure of flavin oxidoreductase (PDB code1qfj). The figure ion the left indicates the domains classified by CATH. 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". The figure on the right shows the eight regions divided by probability. Dark blue, greenish yellow, light blue, gray, pink, orange, and red represent regions, I, II, III, IV, V, VI, and VII, , respectively.

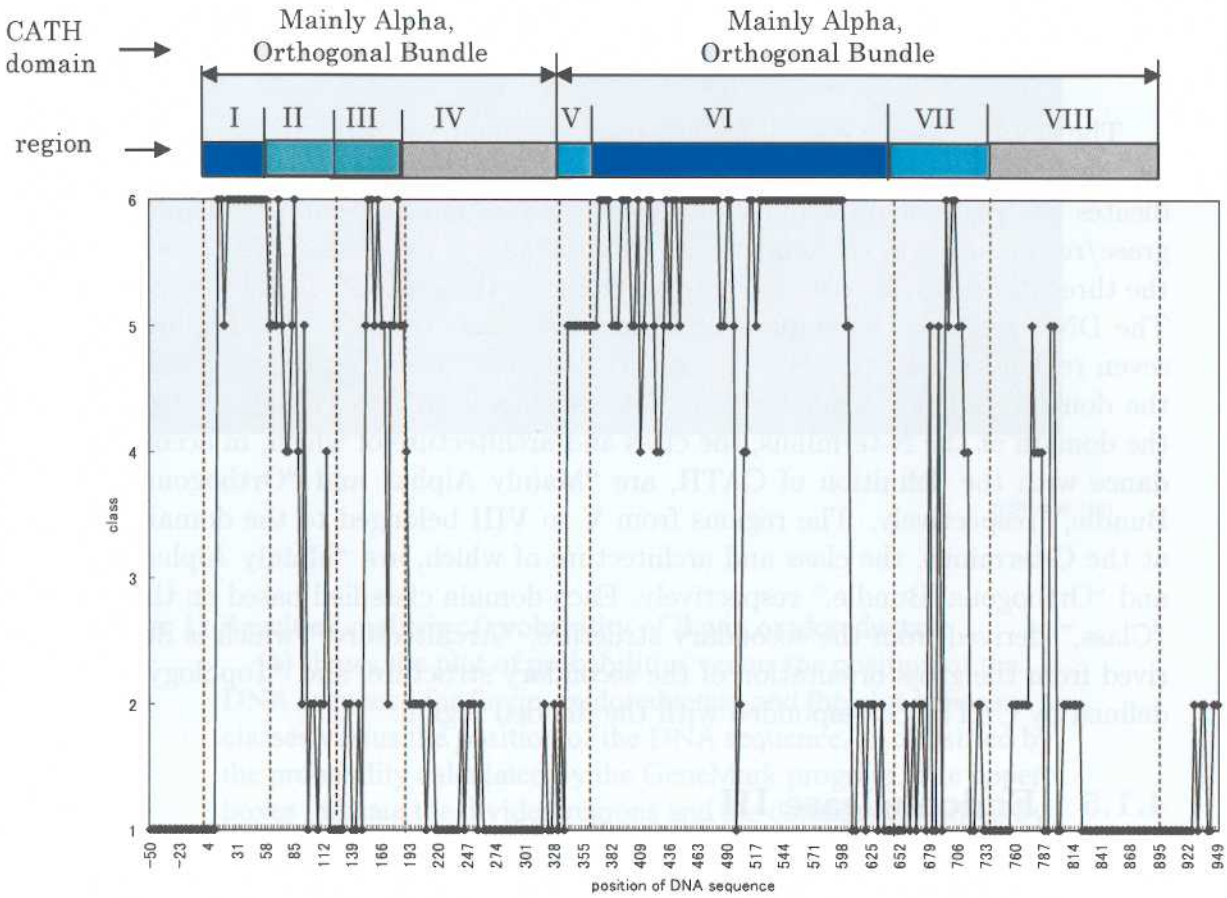
4.1.4 Integrase/recombinase *xer*D

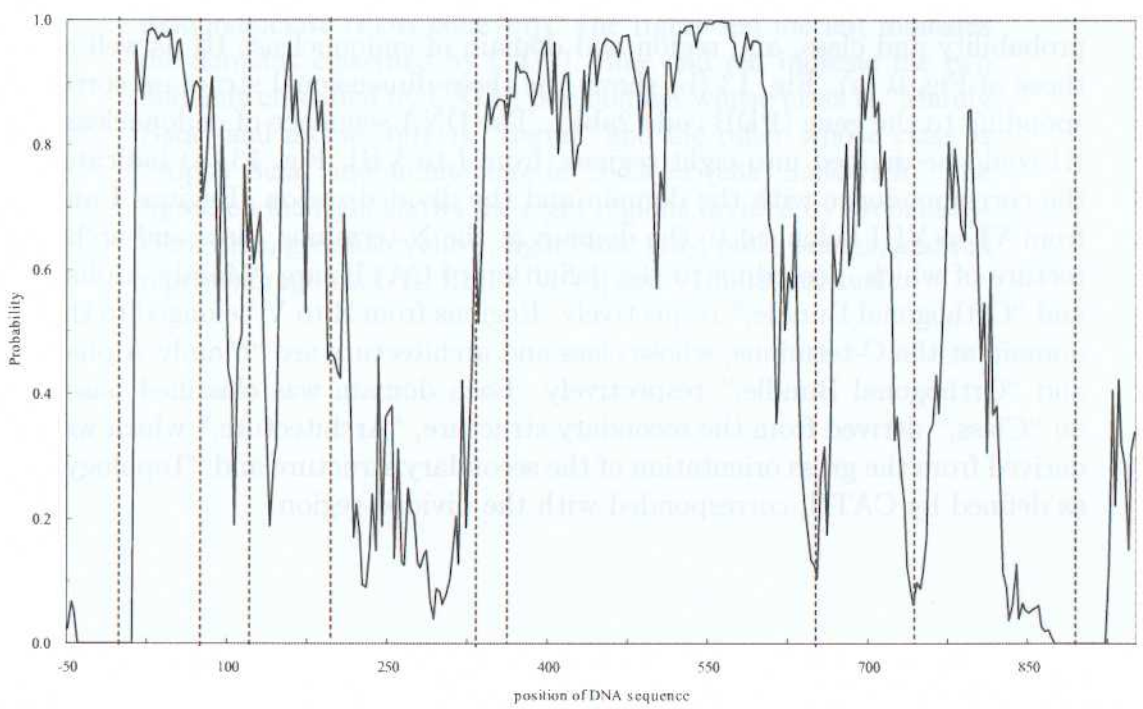
The DNA sequence coding for integrase/recombinase *xer*D was analyzed by the same method as that used for flavodoxin reductase. Fig. 12 (a) indicates the plots of probability and class, and region and domain of integrase/recombinase *xer*D similarly as those of Fig. 9 (a). Fig 12 (b) shows the three-dimensional 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 VIII. Fig. 12 (a) indicates the correspondence with the domain and the divided region. The regions from I to IV 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 V to VIII 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.

4.1.5 Endonuclease III

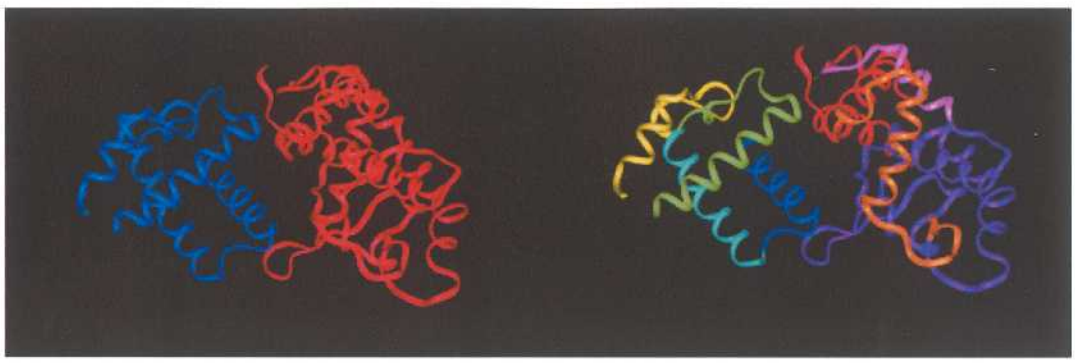
The DNA sequence coding for endonuclease III was analyzed by the same method as that for flavodoxin reductase. Fig. 13 (a) indicates the plots of probability and class, and region and domain of endonuclease III as well as those of Fig. 9 (a). Fig. 13 (b) shows the three-dimensional structure corresponding to the gene (PDB code 2abk). The DNA sequence of endonuclease III could be divided into eight regions, from I to VIII. Fig. 13 (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) The three-dimensional structure of integrase/recombinase *xer*D



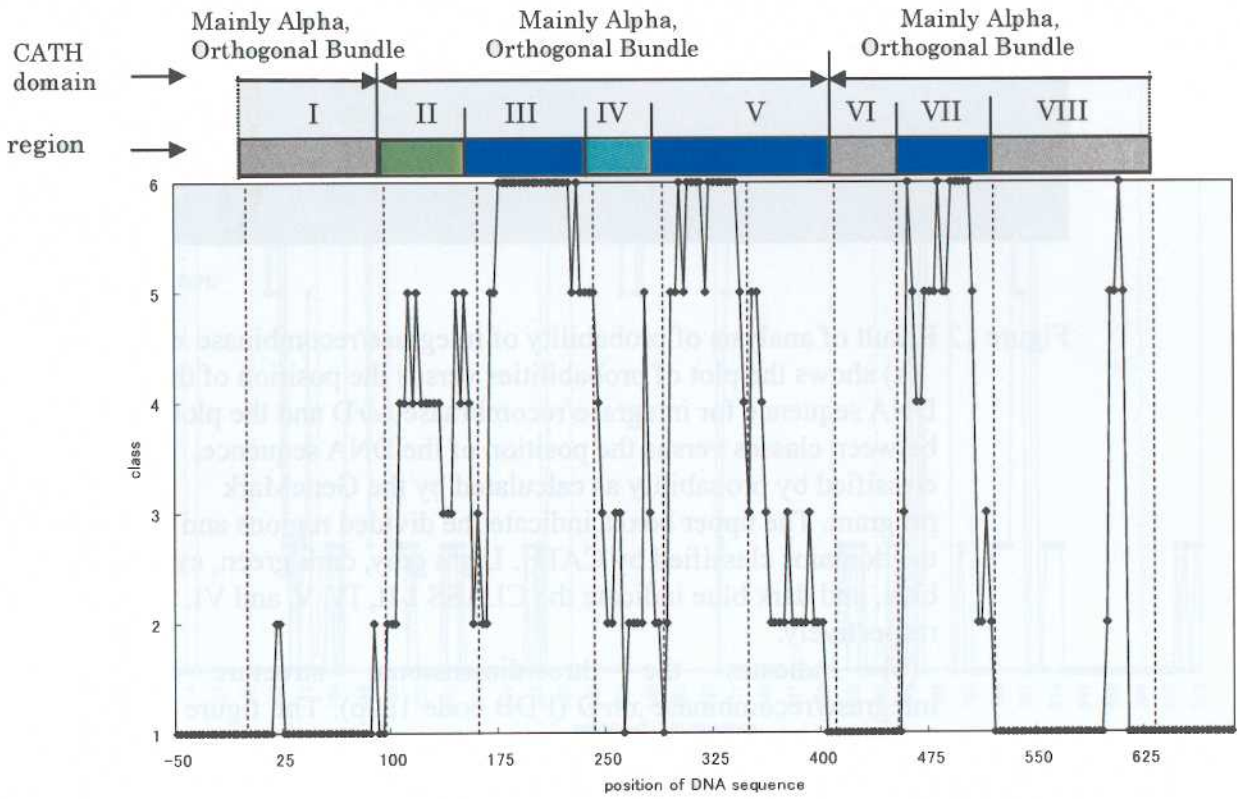
(PDB code: 1a0p)

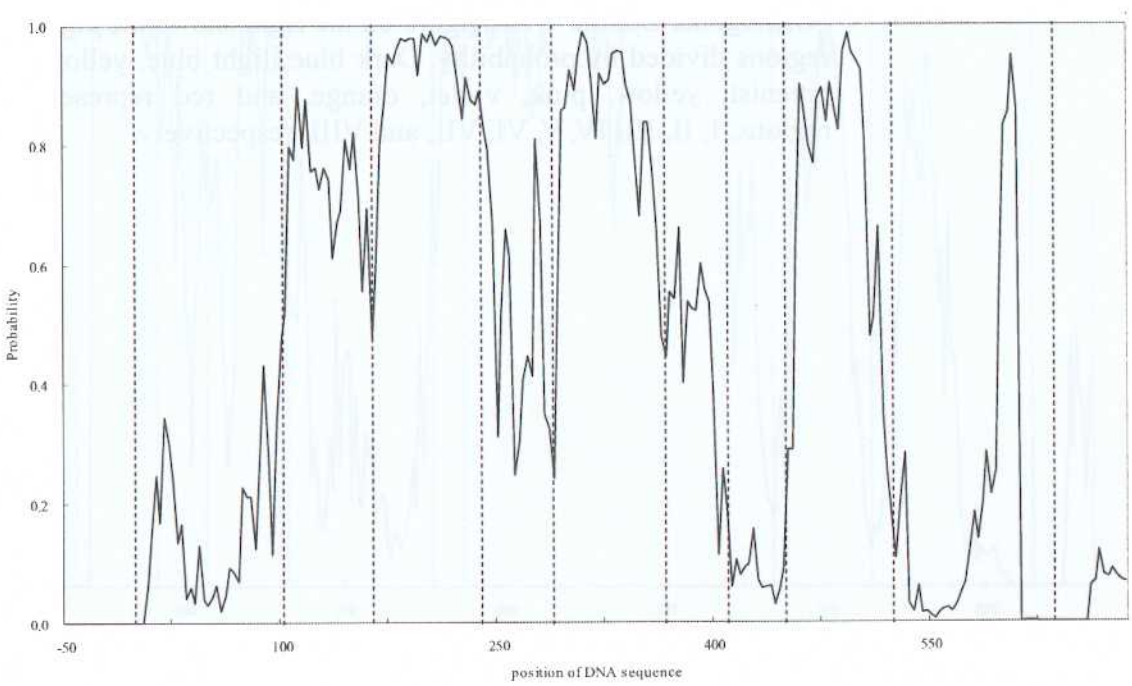
Figure 12 Result of analysis of probability of integrase/recombinase *xer*D

(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 the CLASS I-II, IV, V, and VI, respectively.

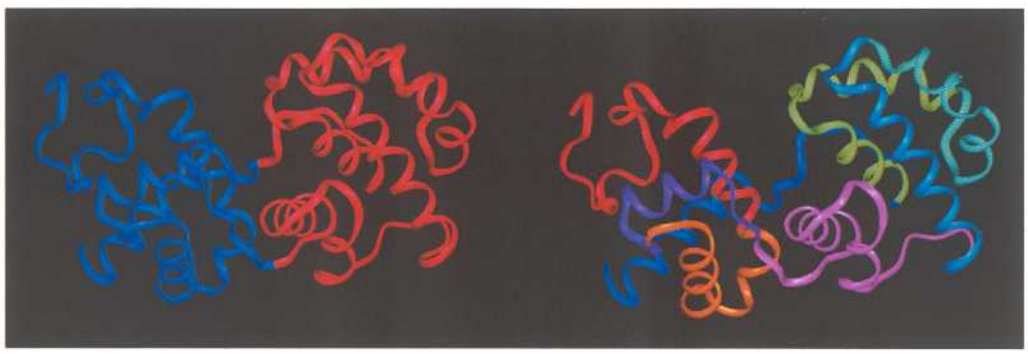
(b) indicates the three-dimensional structure of integrase/recombinase *xer*D (PDB code 1a0p). The figure on the left indicates the domains classified by CATH. Blue and red indicate the two domains classified by CATH; the class and architecture of both domains are "Mainly Alpha" and "Orthogonal Bundle". The figure on the right shows the eight regions divided by probability. Dark blue, light blue, yellow, greenish yellow, pink, violet, orange, and red represent regions, I, II, III, IV, V, VI, VII, and VIII, respectively.

(a) Regions divided by six classes of probability





(b) The three-dimensional structure of endonuclease III



(PDB code 2abk)

Figure 13 Result of analysis of probability of endonuclease III

(a) shows the plot of probabilities versus the position of the DNA sequence of 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 the CLASS I-II, IV, V, and VI, respectively.

(b) indicates the three-dimensional structure of endonuclease III (PDB code 2abk). The figure on the left indicates the domains classified by CATH. Blue and red indicate the two domains classified by CATH, the class and architecture of both domains are "Mainly Alpha" and "Orthogonal Bundle". The figure on the right shows the eight regions divided by probability. Dark blue, greenish yellow, light blue, blue, pink, violet, orange, and red represent regions, I, II, III, IV, V, VI,VII and VIII, respectively.

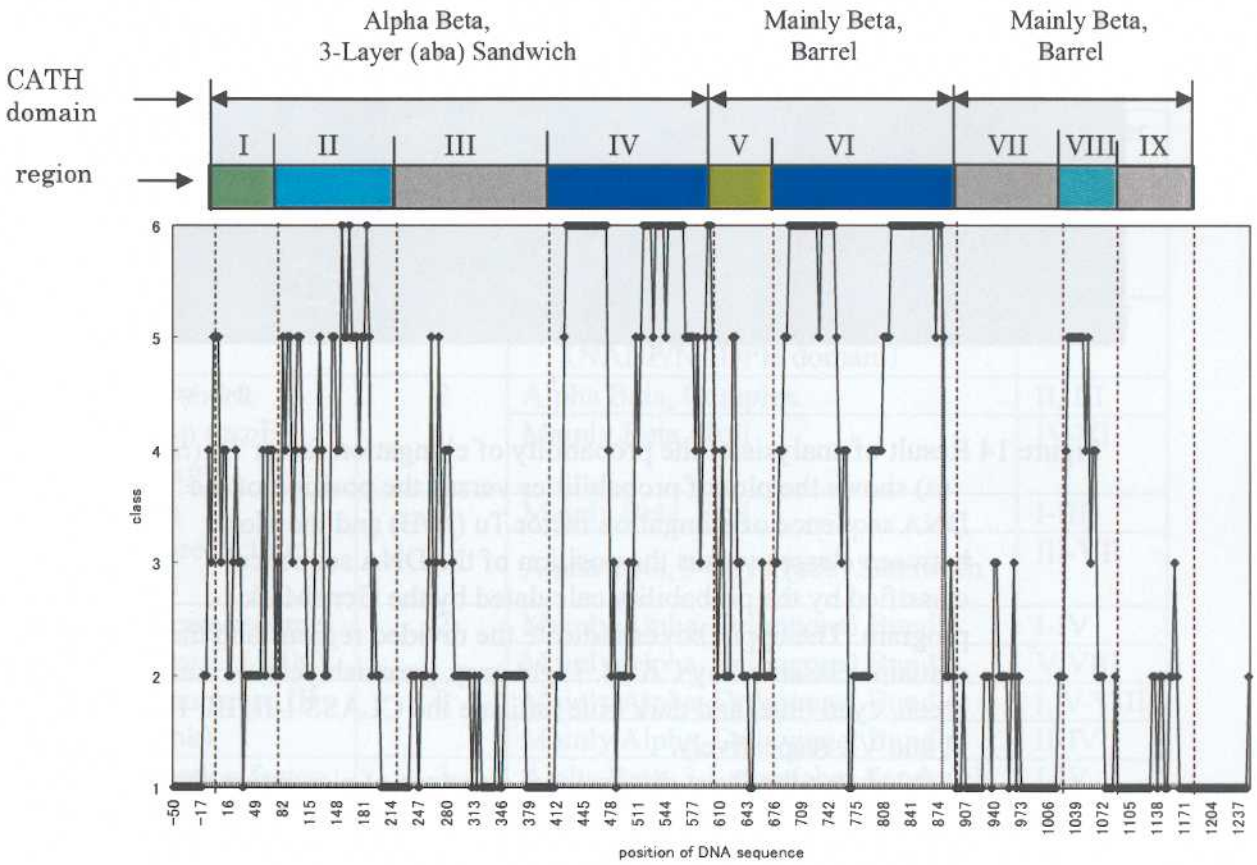
4.1.6 Elongation factor Tu (*tuf*B)

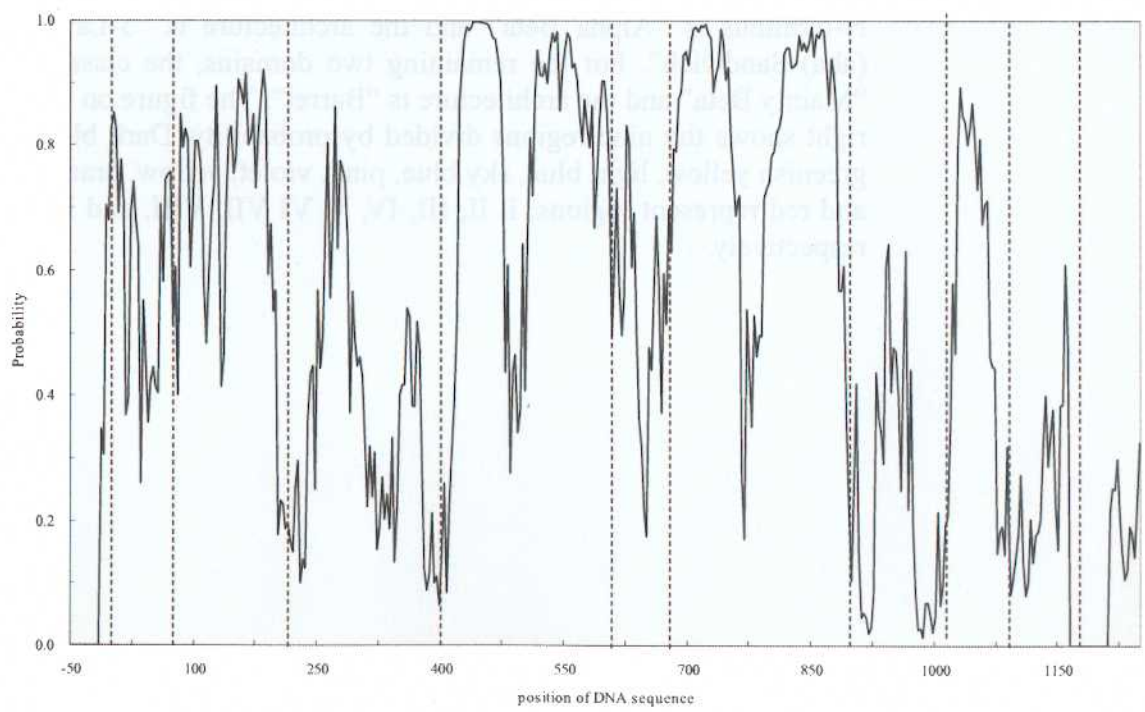
The DNA sequence of elongation factor Tu (*tuf*B) was analyzed by the same method as that used for flavodoxin reductase. Fig. 14 (a) indicates the plots of probability and class, and region and domain of elongation factor Tu (*tuf*B) similarly as those of Fig. 9 (a). Fig. 14 (b) shows the three-dimensional 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 IX. Fig. 14 (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 form VII to IX 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.

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

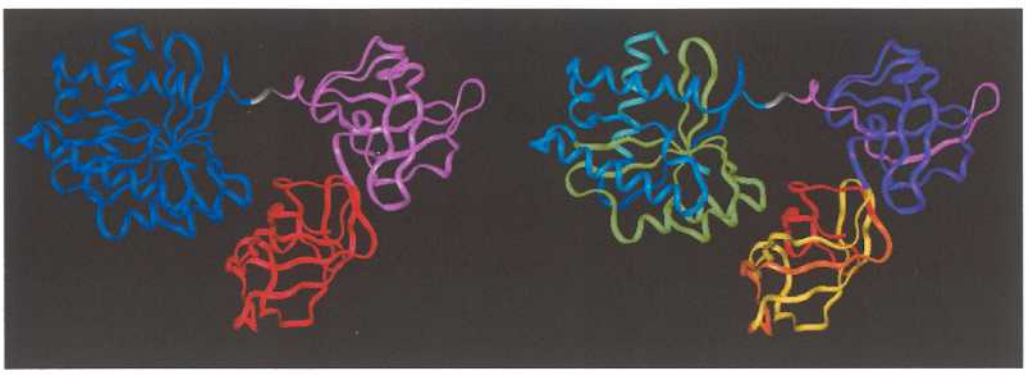
Table 4 indicates the summary of the comparison between the divided regions and the domain classified in accordance with the definition of CATH. The divided regions of five genes, except for flavin oxidoreductase in *Escherichia coli* by the probability calculated using Markov models, corresponded to 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.

(a) Regions divided by six classes of probability





(b) The three-dimensional structure of elongation factor Tu



(pdb code 1efu)

Figure 14 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 the CLASS I-II, III, IV, V, and VI, respectively.
(b) shows the three-dimensional structure of elongation factor Tu (PDB code 1efu). The figure on the left indicates the domains classified by CATH. 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". The figure on the right shows the nine regions divided by probability. Dark blue, greenish yellow, light blue, sky blue, pink, violet, yellow, orange, and red represent regions, I, II, III, IV, V, VI VII, VIII, and IX, respectively.

Table 4 Comparison between the divided regions and the domains classified in accordance with the definitions of CATH

Gene (PDB code)	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-VIII
heat shock protein (*grp*E) (1dkg)	2	Alpha Beta, Complex	II, III
		Mainly Beta, Roll	IV-VI
flavin oxidoreductase (1qfj)	2	Mainly Beta, Roll	I-III
		Alpha Beta, 3-Layer(aba) Sandwich	III-VII
integrase/recombinase *xer*D (1a0p)	2	Mainly Alpha, Orthogonal Bundle	I-IV
		Mainly Alpha, Orthogonal Bundle	V-VIII
endonuclease III (2abk)	2	Mainly Alpha, Orthogonal Bundle	I, V-VIII
		Mainly Alpha, Orthogonal Bundle	II-IV
elongation factor Tu (*tuf*B) (1efu)	3	Alpha Beta, 3-Layer (aba) Sandwich	I-IV
		Mainly Beta, Barrel	V, VI
		Mainly Beta, Barrel	VII-IX