Chapter 5

Structural evolution of protein

5.1 Flavodoxin reductase of *Escherichia coli*

5.1.1 DNA sequence, amino acid sequence, three-dimensional structure and the divided regions

Fig. 15 shows the DNA sequence [48], the amino acid sequence [42], the secondary structures [42], and the divided regions [59] of flavodoxin reductase in *Escherichia coli*. Fig. 16 shows the plot of class against the position of the DNA sequence of flavodoxin reductase, calculated with Markov models by the GeneMark program. Fig. 17 represents the three-dimensional structure of flavodoxin reductase (PDB code 1fdr).

5.1.2 Homology search with the DNA sequences of the divided regions

(1) Regions I and II

If the region, which is adjacent to another region of the same CLASS, the two regions are considered to be a single region because of being derived from the same species. Thus, a homology search was done with the DNA sequence of region I+II.

The DNA sequence of region I+II was homologous to those of NADPH:ferredoxin reductase from *Xanthomonas campestris* pv. *campestris* [60], *Azotobacter vinelanndii* [61], *Pseudomonas aeruginosa* PA01 [62], or *Xylella fastidiosa* 9a5c [63], of the soluble methane monooxygenase gene from *Methylomonas* sp. KSWIII [63], and protein-glutamate methylesterase from *Ralstonia solanacearum*

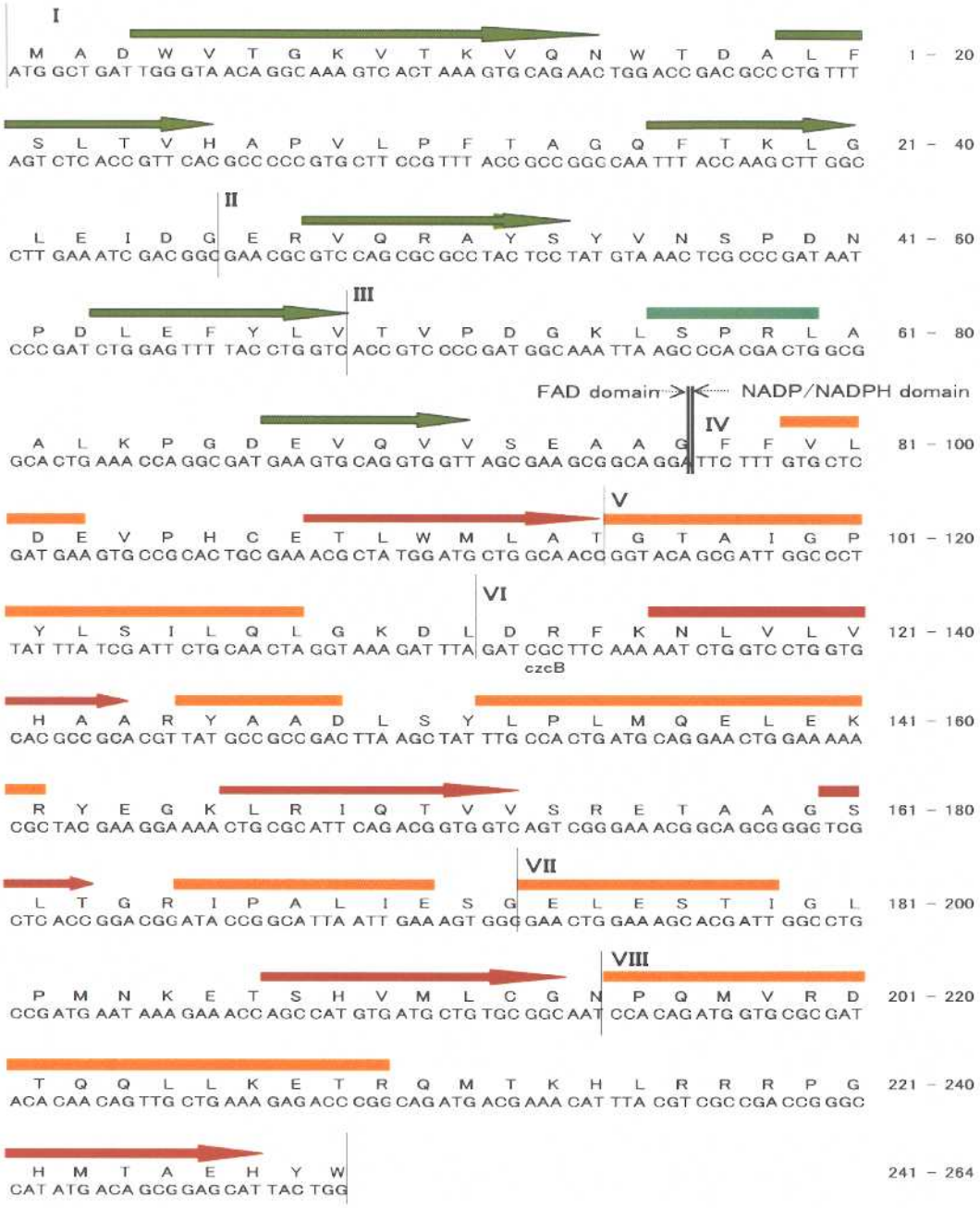


Figure 15 DNA sequence, amino acid sequence, secondary structures and the divided regions of flavodoxin reductase form *Escherichia coli*

A light green box and dark green arrows indicate the α-helix and β-sheet structures of the FAD domain. Orange boxes and brown arrows indicate α-helices and β-sheets of the NADP domain. Characters from I to VIII show the region divided by class.

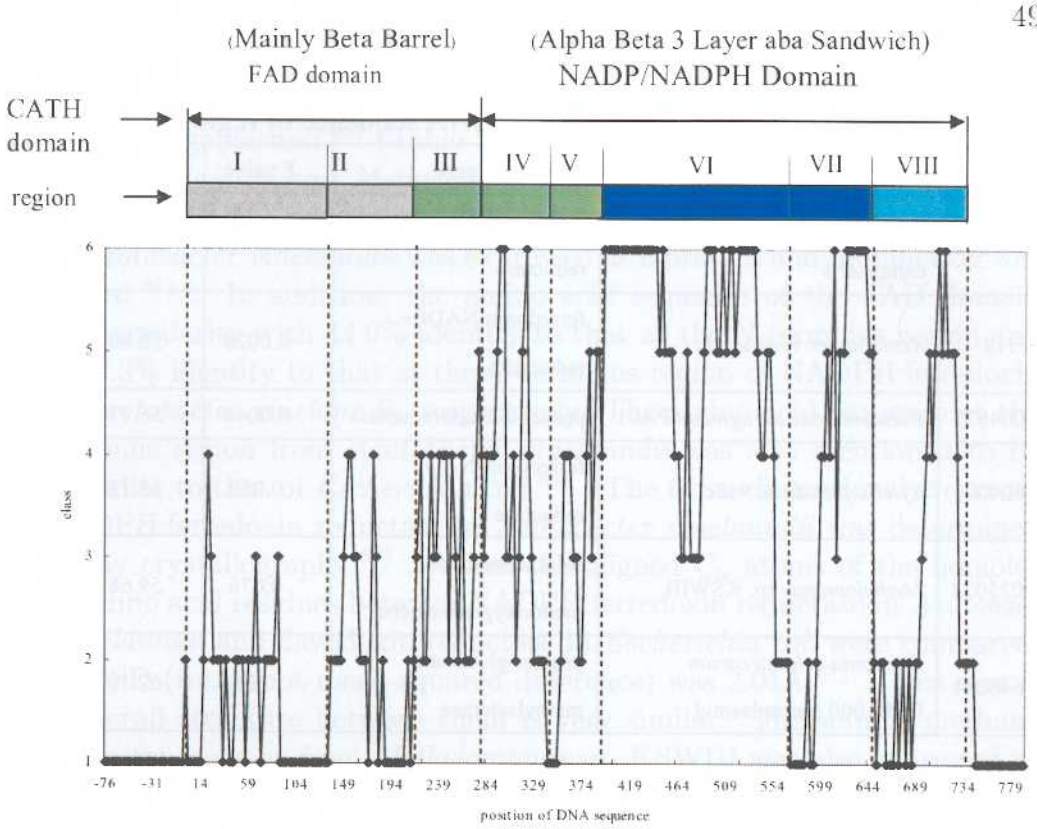
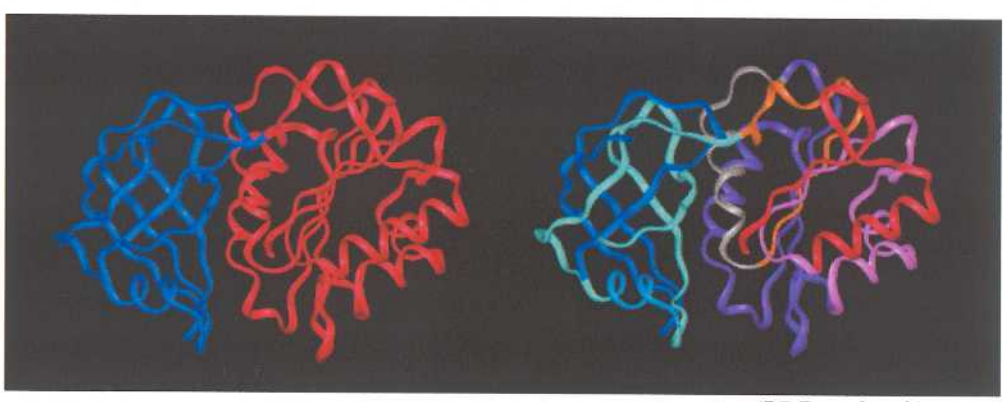


Figure 16 Regions divided by probability of flavodoxin reductase

The regions divided by six classes of probability were calculated by the GeneMark program. Light gray, dark green, light blue and dark blue boxes indicate the CLASS I-II, IV, V, and VI, respectively.



(PDB code 1fdr)

Figure 17 The three-dimensional structure of flavodoxin reductase

The figure on the left indicated the domains classified by CATH. Dark blue and red ribbons represent the FAD domain and the NADP domain, respectively. 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.

Table 5 Result of FASTA search with the DNA sequence of region I+II

GI	species	gene	E value	identity
AE012241	*Xanthomonas campestris* pv. campestris	ferredoxin-NADP reductase	0.00078	58.85%
L36319	*Azotobacter vinelanndii*	ferredoxin NADP+ reductase	0.0028	55.80%
AE004875	*Pseudomonas aeruginosa* PA01	probable oxidoreductase	0.0044	56.77%
AE004009	*Xylella fastidiosa* 9a5c	ferredoxin-NADP reductase	0.059	55.73%
AB025022	*Methylomonas* sp. KSWIII	soluble methane monooxygenase gene	0.076	59.68%
AL646084	*Ralstonia solanacearum* GMI1000 megaplasmid	protein-glutamate methylesterase	0.31	62.19%

GMI1000 megaplasmid [65] (Table 5). The genes except for those from *Azotobacter vinelanndii* and *Methylomonas* sp. KSWIII were predicted on the basis of their DNA sequences. In contrast, NADPH:ferredoxin reductase from *Azotobacter vinelanndii* was expressed as a protein and its function was identified [61] . In addition, 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 amino acid sequence at the N-terminus region from *Azotobacter vinelanndii* was also mentioned to be very similar to that of *Escherichia coli* [61] . The three-dimensional structure of NADPH:ferredoxin reductase in *Azotobacter vinelanndii* was determined by X-ray crystallography [66] . When the aligned C_α atoms of the homologous amino acid residues between NADPH:ferredoxin reductase in *Azotobacter vinelanndii* and flavodoxin reductase in *Escherichia coli* were compared, the RMSD (total root-mean-squared difference) was 2.01Å [67] . This means that overall structure between them is very similar. The soluble methane monooxygenase gene from *Methylomonas* sp. KSWIII was also expressed as a protein and its function was identified [64] . But NADPH:ferredoxin reductase from *Azotobacter vinelanndii* showed higher similarity to the DNA sequence of region I+II than did the soluble methane monooxygenase gene of *Methylomonas* sp. KSWIII. The DNA sequence of region I+II, where FAD binds, would have been come from NADPH:ferredoxin reductase in *Azotobacter vinelanndii.*

(2) Region III

The CLASS of the region III is the same as that of the region IV. But the classes of probability in the rtegion III are those from 2 to 4. On the contrast, the classes of the probability in the region IV are those from 2 to 6. The region III is belonging to the FAD domain and the region IV the NADP domain, respectively. Because the pattern of the class between them is different and each region locates at the different domains determined by X-rayt chrystallography, I did not combine them. I did homology serch with the region III. No gene, whose E-value was less than 1.0, was found to be homologous to the DNA sequence of region III.

(3) Region IV and V

I did homlogy search with the region IV+V, because the pattern of the classes is similar and the CLASS of the region is the same between them. The DNA sequence of region IV+V shows homology to that of lytB (involved in polymyxin B resistance) from *Burkholderia pseudomallei* [68] , the conjugal transfer gene E (*tra*E) from *Escherichia coli* pKM101 [69] or IncN plasmid R46 [70] , the hypothetical protein MmyD from *Streptomyces coelicolor* plasmid SCP1 [71] and an unidentified sequence from *Hydrogenophaga pseudoflava* [72] or *Sphingomonas paucimobilis* [73] (Table 6). It was noted that *Escherichia coli* pKM101 was derived by deletion of the region related to heavy metal and drug resistance from IncN plasmid R46, by comparison with maps of restriction enzyme between *Escherichia coli* pKM101 and IncN plasmid R46 [74] . Genes coded on a plasmid can be transferred via cell-to-cell conjugation from one species to another. Conjugal transfer gene E (*tra*E) is one of the genes located in the *tra* operon which is expressed through conjugation. Because the class of this region is greater than those of the regions I, II and III, the DNA sequence of this region would be coming from horizontally transferred genes, which were used for the Markov model made by Borodovsky M. *et al.*. Therefore, the DNA sequence of region IV, where NADP/NADPH binds, would have been transferred from the *Escherichia coli* plasmid pKM101.

(4) Regions VI and VII

Since regions VI and VII show the same CLASS and the same pattern of the class, the two regions were regarded as a single region. Therefore the homology search was done for region VI+VII. The DNA sequence of region VI+VII was homologous to that of *czc*B (membrane fusion protein) from *Ralstonia* sp. CH34 pMOL30 [75] or *Alcaligenes* sp. [76] (Table 7). Since the class of region VI+VII is high, it is considered likely that the DNA sequence of this region would have come from the horizontally transferred genes. Consequently, the DNA sequence of region VI+VII, where NADP/NADPH binds, would have been derived from the *Ralstonia* sp. CH3 pMOL30. The *czc*B gene from the *Ralstonia* sp. CH34 pMOL30 may function as a cation binding subunit and act as a funnel in the mechanism of the cation-proton antipoter [77] .

Table 6 Result of FASTA search with the DNA sequence of region IV+V

GI	species	gene	E value	identity
D63999	*Synechocystis* sp. PCC 6803	*sph*S	0.36	70.46%
AF009224	*Acinetobacter* sp. ADP1	*ben*C, reductive oxygenase for benzoate	0.57	72.50%
AF314502	*Bacillus subtilis*	glucose dehydrogenase	0.81	80.56%

Table 7 Result of FASTA search with the DNA sequence of region VI+VII

GI	species	gene	E value	identity
X98451	*Ralstonia metallidurans* CH34 plasmid pMOL30	*czc*B, membrane fusion protein	0.0068	64.42%
D67044	*Alcaligenes* sp.	*czc*B, membrane fusion protein	0.029	63.46%

(5) Region VIII

The DNA sequence of region VIII was homologous to that of DNA repair exonuclease (*sbc*D) from *Bacillus subtilis* [78] , *rhc*U from *Bradyrhizobium japonicum* [79] or orf17 from *Streptococcus pneumoniae* bacteriophage Cp-1 [80] (Table 8). These genes except for that from *Streptococcus pneumoniae* bacteriophage Cp-1 were predicted based on the DNA sequences. The orf17 gene homologous to the DNA sequence of this region was expressed as a protein and deduced to be the tail protein by the high similarity to the N-terminal amino acid sequence from bacteriophage ϕ29 [80] . Since the class of this region is high, it is considered likely that the DNA sequence of this region might have come from horizontally transferred genes. Therefore, the DNA sequence of region VII, where NADP/NADPH binds, would have been come from orf17 in the *Streptococcus pneumoniae* bacteriophage Cp-1.

5.2 Heat shock protein (*grp*E) of *Escherichia coli*

5.2.1 DNA sequence, amino acid sequence, three-dimensional structure and the divided regions

Fig. 18 shows the DNA sequence [48] , the amino acid sequence [43] , the secondary structures [43] and the divided regions [59] of heat shock protein (*grp*E) in *Escherichia coli*. Fig. 19 shows the plot of class against position of the DNA sequence of heat shock protein (*grp*E), calculated with Markov models by the GeneMark program. Fig. 20 represents the three-dimensional structure of heat shock protein (PDB code 1dgk).

5.2.2 Homology search with the DNA sequences of the divided regions

If the adjacent regions are the same CLASS, they were considered as a single region. Thus homology search was done with the DNA sequence of region II+III. As the region I and the region IV are extremely short DNA sequences for homology search, the region I and the region IV were combined with the region II+III and the region V, respectively.

Table 8 Result of FASTA search with the DNA sequence of region VIII

GI	species	gene	E value	identity
Y09476	*Bacillus subtilis*	DNA repair exonuclease (*sbc*D)	0.087	60.53%
AF322012	*Bradyrhizobium japonicum*	*rhc*U	0.1	63.95%
Z47794	*Streptococcus pneumoniae* Bacteriophage Cp-1	orf17	0.32	62.00%

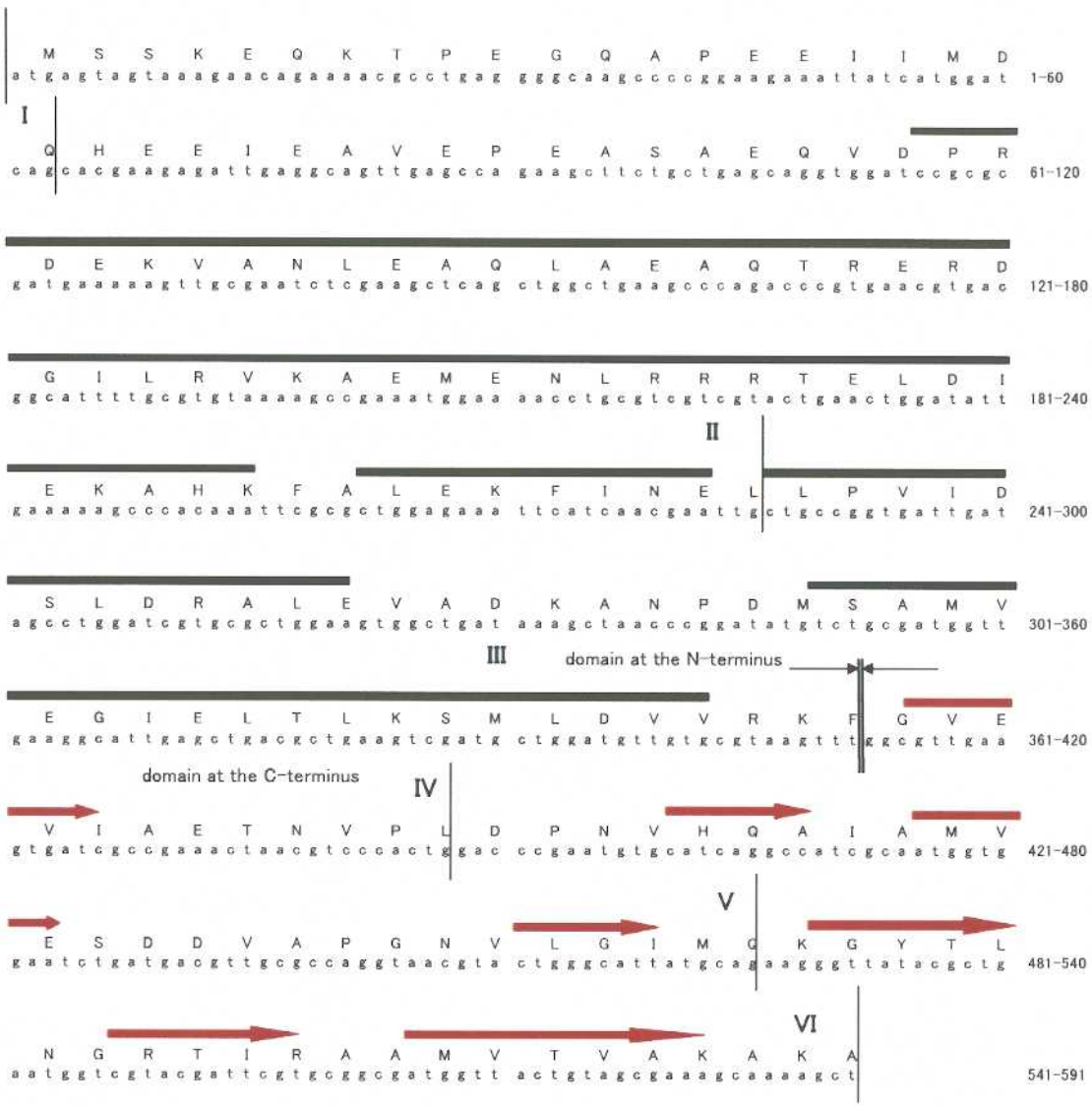


Figure 18 DNA sequence, amino acid sequence, secondary structures and the divided regions of heat shock protein (*grp*E) from *Escherichia coli*

Green boxes and brown arrows indicate the α-helix of the domain at the N-terminus determined by X-ray crystallography and the β-sheet of the domain at the C-terminus, respectively. Characters from I to VI represent the regions divided by class. The three-dimensional structure of the amino aid sequence from position 1 to 33 did not have been determined.

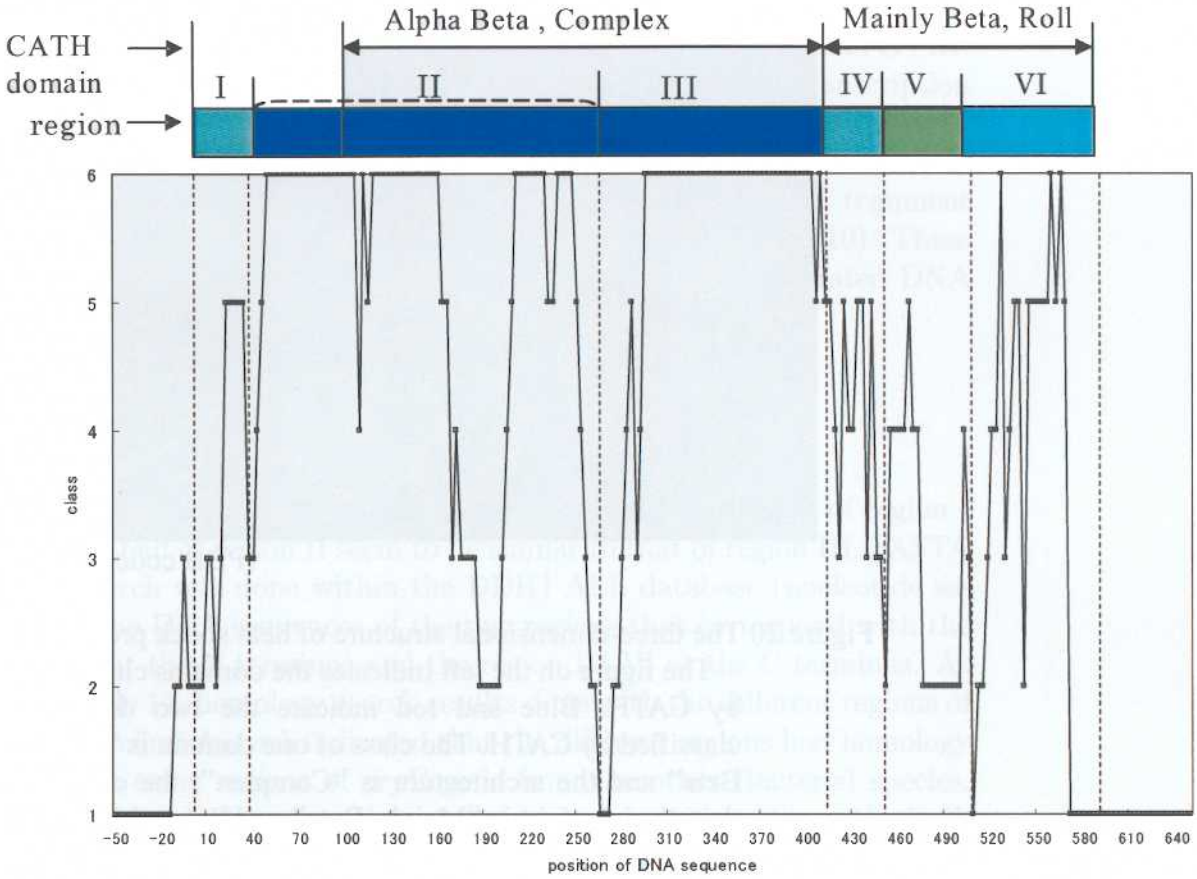


Figure 19 Regions divided by probability of heat shock protein (*grp*E)
The region divided by six classes of probability calculated by program GeneMark. Dark green, light blue and dark blue boxes indicate the CLASS VI, V, and VI, respectively.

(PDB code: 1dkg)

Figure 20 The three-dimensional structure of heat shock protein (*grp*E)
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. Blue, sky blue, pink, orange, and red represent a part of region II, region III, region IV, region V, and region VI, respectively

No DNA sequence was found to be homologous to that of region I+II+III. The region IV+V showed homology to the DNA sequence of transcription antitermination protein (*nus*G) from *Streptomyces coelicolor* [81] (Table 9). The region VI had homology to the DNA sequence of heat shock protein (*grp*E) from *Nitrosomonas europaea* [82] , and that of very large tegument protein (UL36) from *Gallid herpesvirus* 1 (serotype 2) [83] (Table 10). These results do not seem to make a good scenario of having incorporated DNA sequence as parts of structure.

5.2.3 Duplication of gene

By looking at the pattern of class of probability, the pattern of region I and the front half of region II seem to be similar to that of region III. FASTA homology search was done within the DDBJ ALL database (nucleotide sequence) for the DNA sequences of the two regions that correspond with the region I +II at the N-terminus and the region III-VI at the C-terminus. As shown in Table 11, homology search results done with the different regions of *grp*E from *Escherichia coli* indicated that the different regions had homology to almost the same regions of *grp*E gene from the other Bacterial species, but not from any Eukarya species or any Archaeal species. Thus, the both DNA sequences of region I + II and region III-VI of the *grp*E gene from *Escherichia coli* were aligned by using the CLUSTAL W program. Fig. 21 represented the alignment of the two regions. The score of nucleotide identity between the two regions was approximately 47%. Accordingly, the heat shock protein (*grp*E) might have been created by fusion after duplication of almost a half of the DNA sequence of *grp*E.

Moreover, the amino acid sequences of Heat shock protein (GrpE) from 7 species, such as *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 of the MED1 (the GrpE homolog) from *Saccharomyces cerevisiae*, mitochondoria of Eukarya were homologous to that of the Heat shock protein (GrpE) from *Escherichia coli*. These 10 GrpE proteins were aligned by using the CLUSTAL W program. As shown in Fig. 22, the alignment derived from 10 species demonstrated that amino acid sequences at the C-terminus had higher homology than those at the N-terminus. It can be suggested that the *grp*E gene might have evolved by duplication of the DNA sequence at the C-terminus followed by fusion of the two fragments. And the gene before duplication seems to have already disappeared.

Table 9 Results of FASTA search with the DNA sequence of region IV+V

GI	species	gene	E-value	identity
AL160431	*Streptomyces coelicolor*	transcription antitermination protein (*nus*G)	0.62	62.69%

Table 10 Results of FASTA search with the DNA sequence of region VI

GI	species	gene	E-value	identity
AB018706	*Nitrosomonas europaea*	heat shock protein (*grp*E)	0.0044	68.25%
AB024309	*Gallid herpesvirus* 1 (serotype 2)	very large tegument protein (UL36)	0.089	70.83%

Table 11 FASTA homology search results within DDBJALL (nucleotide sequence) with the DNA sequences of region I+II and region III-VI of the *grp*E gene from *Escherichia coli*

region	accession number	*species*	E value	identity	homologous region of query sequence	coresponding region of homologous gene
I+II	X07863	*Escherichia coli grpE*	2.70E-76	—	1-303	616-918
III-VI	X07863	*Escherichia coli grpE*	4.10E-82	—	1-303	718-1020
I+II	AL627276	*Salmonella enterica* serovar Typhi *grpE*	5.40E-61	89.44%	303-1	35784-36086
III-VI	AL627276	*Salmonella enterica* serovar Typhi *grpE*	2.10E-61	86.80%	303-1	35682-35984
I+II	AE008821	*Salmonella typhimurium* LT2 *grpE*	9.10E-61	89.44%	303-1	20018-20320
III-VI	AE008821	*Salmonella typhimurium* LT2 *grpE*	3.50E-61	86.80%	303-1	19916-20218
I+II	AJ414146	*Yersinia pestis* strain CO92 *grpE*	6.40E-15	63.22%	47-301	131738-131997
III-VI	AJ414146	*Yersinia pestis* strain CO92 *grpE*	8.70E-28	68.95%	28-303	131826-132101
I+II	AF218211	*Vibrio proteolyticus* GrpE *(grpE)*	3.60E-10	72.22%	178-303	354-479
III-VI	AF218211	*Vibrio proteolyticus* GrpE *(grpE)*	5.30E-17	66.97%	76-294	354-572
I+II	AE004890	*Pseudomonas aeruginosa PA01 grpE*	4.80E-07	68.97%	303-188	397-512
III-VI	AE004890	*Pseudomonas aeruginosa PA01 grpE*	8.00E-15	65.48%	282-86	316-512
I+II	U32693	*Haemophilus influenzae* Rd *grpE*	7.40E-07	60.31%	114-303	289-481
III-VI	U32693	*Haemophilus influenzae* Rd *grpE*	2.60E-10	61.43%	12-217	289-497
I+II	AF106835	*Methylovorus sp.* strain SS1 *grpE*	4.00E-07	69.57%	188-302	295-409
III-VI	AF106835	*Methylovorus sp.* strain SS1 *grpE*	1.00E-10	66.67%	86-244	295-450
I+II	AE004170	*Vibrio cholerae* chromosome I *grpE*	2.60E-06	65.87%	178-303	8554-8679
III-VI	AE004170	*Vibrio cholerae* chromosome I *grpE*	2.90E-12	62.16%	76-294	8554-8772
I+II	AE006069	*Pasteurella multocida* PM70 *grpE*	1.50E-05	60.48%	135-301	3824-3989
III-VI	AE006069	*Pasteurella multocida* PM70 *grpE*	1.70E-07	60.56%	33-212	3824-4002
I+II	AE011784	*Xanthomonas axonopodis* pv. citri *grpE*	1.40E-04	64.66%	188-303	1356-1471
III-VI	AE011784	*Xanthomonas axonopodis* pv. citri *grpE*	9.80E-07	66.12%	86-206	1356-1476
I+II	L43367	*Francisella tularensis grpE*	3.50E-04	64.10%	186-302	400-516
III-VI	L43367	*Francisella tularensis grpE*	1.40E-07	60.18%	84-303	400-613
I+II	AF302775	*Xanthomonas campestris* pv.	5.80E-04	63.79%	188-303	1454-1569
III-VI	AF302775	*Xanthomonas campestris pv.*	3.80E-06	65.12%	86-214	1454-1581
I+II	AE012247	*Xanthomonas campestris pv.*	4.60E-04	63.79%	188-303	12346-12461
III-VI	AE012247	*Xanthomonas campestris pv.*	3.00E-06	65.12%	86-214	12346-12473
I+II	X51477	*Bacillus subtilis grpE*	7.60E-02	62.19%	185-302	441-558
III-VI	X51477	*Bacillus subtilis grpE*	2.30E-04	61.97%	83-223	441-581

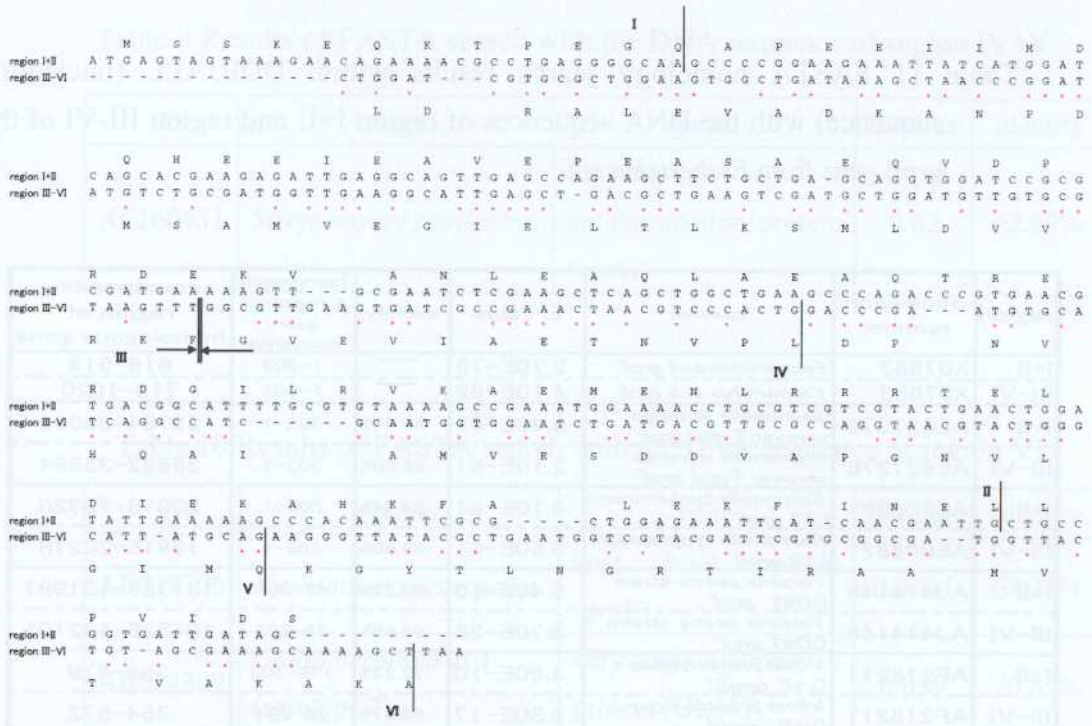


Figure 21 The alignment of the region I +II and the regions III-VI of the *grp*E from *Escherichia coli*

The alignment of the DNA sequences was produced by using the CLUSTAL W program. The first and fifth raw indicate the amino acids corresponding to the DNA sequence of region I+II of the *grp*E at the N-terminus and that of region III-VI of the *grp*E at the C-terminus, respectively. The second and third raw indicate the DNA sequence of region I+II of the grpE at the N-terminus and that of regions III-VI of the grpE at the C-terminus, respectively. The asterisks of the fourth raw indicate the same nucleotides in the alignment.

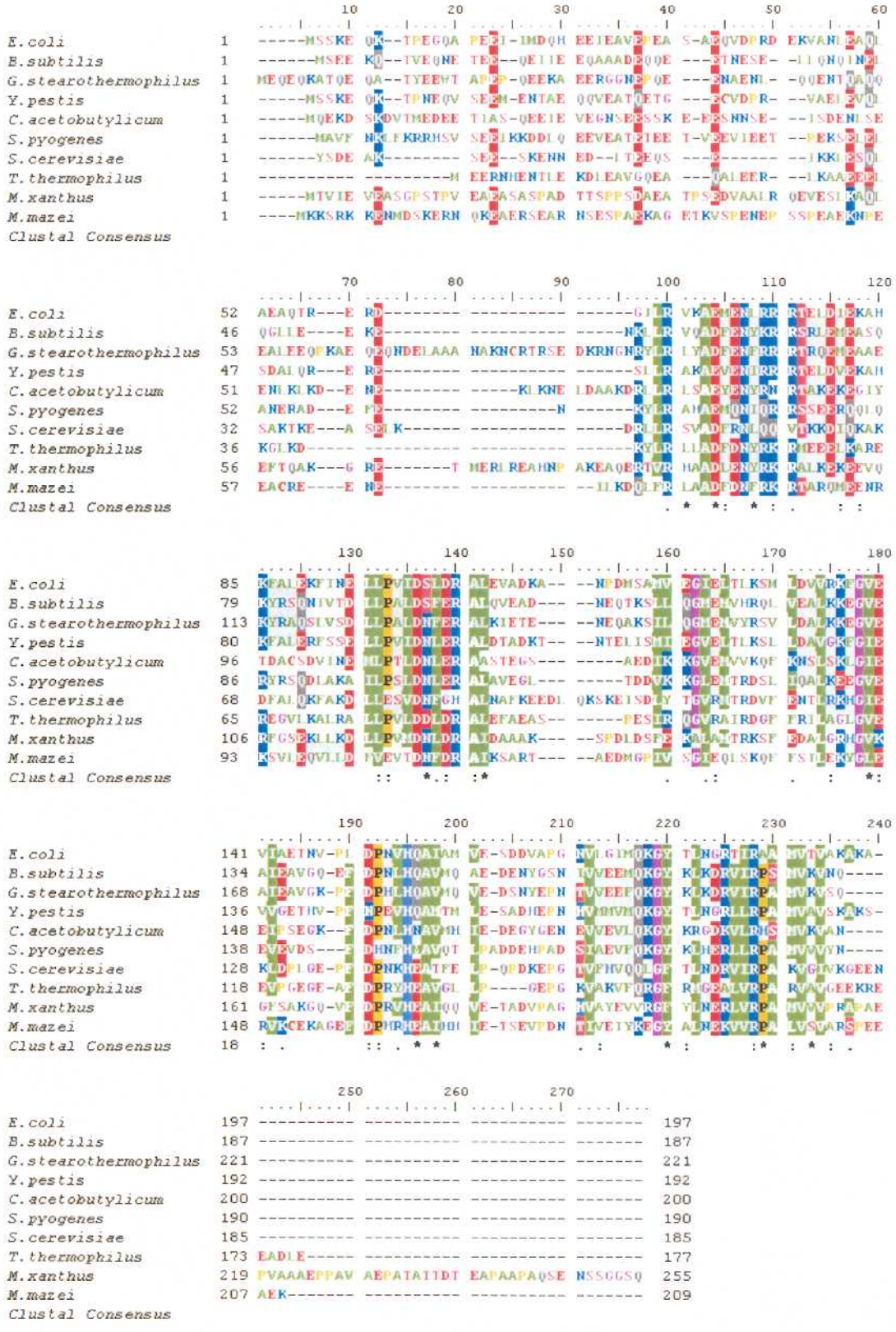

E.coli
B.subtilis
G.stearothermophilus
Y.pestis
C.acetobutylicum
S.pyogenes
S.cerevisiae
T.thermophilus
M.xanthus
M.mazei
Clustal Consensus

Figure 22 The alignment of amino acid sequences for Heat shock protein (GrpE) from 10 species.

The amino acid sequences from *Escherichia coli*, *Bacillus subtilis*, *Clostridium acetobutylicum*, *Streptococcus pyogenes*, *Thermus thermophilus*, *Myxococcus xanthus*, *Methanosarcina mazei*, *Geobacillus stearothermophilus*, *Yersinia pestis*, and *Saccharomyces cerevisiae* were derived from P09372, P15874, P30726, Q99YC8, Q56236, P95333 and P42367 in Swiss Protein Database, CAA62238.1 and CAC89950.1 in EMBL, and S41760 in PIR, respectively. The amino acids column was colored when the similarity among them were more than 20%. The amino acid sequences at the C-terminus are more well-conserved than those at the N-terminus.

The abbreviations of *Escherichia coli*, *Bacillus subtilis*, *Geobacillus stearothermophilus*, *Yersinia pestis*, *Clostridium acetobutylicum*, *Streptococcus pyogenes*, *Saccharomyces cerevisiae*, *Thermus thermophilus*, *Myxococcus xanthus*, and *Methanosarcina mazei*, are given as *E.coli*, *B.subtilis*, *G.stearothermophilus*, *Y.pestis*, *C.acetobutylicum*, *S.pyogenes*, *S.cerevisiae*, *T.thermophilus*, *M.xanthus*, and *M.mazei*, respectively. The alignment was produced by using the CLUSTAL W program. The figure was produced using the BioEdit version 5.0.6 program [84].