Chem-Bio Informatics Journal, Vol. 2, No. 4, pp.137-146(2002)

Structural evolution of Flavodoxin reductase in *Escherichia coli*

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

The three-dimensional structure of Flavodoxin reductase in *Escherichia coli* had been determined by X-ray crystallography and was found to consist of multi-domains. Then the DNA sequence of Flavodoxin reductase was analyzed by using Markov models, and the sequence was divided into seven regions (I-VII), which corresponded to the structural domains classified based on "Class," derived from the secondary structure, "Architecture," which was derived from the gross orientation of the secondary structures and "Topology," as defined by CATH. In addition, the domains defined by CATH are equivalent to the domains identified by analysis of the X-ray crystallography.

A FASTA homology search was done against the DDBJ ALL database with each divided region. Region I+II belongs to the FAD domain, which binds FAD, and its DNA sequence showed homology to that of the FAD binding site of NADPH:ferredoxin reductase of *Azotobacter vinelanndii*. It is reported that the overall 3D structure was very similar to that of Flavodoxin reductase, since the RMSD between these two proteins was 1.49 Å and the amino acid sequences at the N-termini also showed high homology. Thus there is a correspondence between the 3D structures, amino acid sequences, and the divided sequences of both sequences. The region from IV to VII belongs to the NADP domain, which binds to NADP/NADPH. The DNA sequences of regions IV, V+VI and VII showed respective homology to conjugal transfer gene E (*tra*E) of *Escherichia coli* pKM101, the *czc*B gene (cation-proton antipoter) from *Ralstonia* sp. CH34 pMOL30, and orf17 of *Streptococcus pneumoniae* Bacteriophage Cp-1, which was predicted as a tail protein. Thus it is suggested that the NADP domain would have been constructed by the fusion of a tail protein from a bacteriophage, the *czc*B gene and the conjugal transfer gene from a plasmid. Moreover, NADPH:ferredoxin reductase of *Azotobacter vinelanndii,* which showed homology to the DNA sequence in the region of the FAD domain, was coded on the chromosome. On the contrary, genes homologous to the DNA sequence of the region of the NADP domain were coded on a plasmid or bacteriophage. The structurally divided regions of Flavodoxin reductase were supposed to be comprised of several fragments of DNA sequence whose imprinted structural information might be incorporated as parts of structure from other species.

Key Words: structure evolution, Flavodoxin reductase, Markov model

Area of Interest: Bioinformatics and Bio Computing

1. Introduction

Information of a living life-form is imprinted on the DNA sequence of chromosomes. Evolutionary events such as mutation or fusion are imprinted on it, too. As a template, the information of the DNA sequence is transferred into mRNA by gene products related to transcription, and mRNA is translated into amino acids by the products of gene related to translation. Then the amino acid sequence is formed into a 3D structure to express its natural function as a protein. Information related to change of function or creation of new function of a protein had thus been imprinted on the DNA, as well. Therefore, we studied the relationship between the structure of protein and evolution of protein function with respect to new points of information on the DNA sequence.

In our previous paper, we analyzed the pattern of probability at each position of the DNA sequence of a gene by calculations using Markov models. In this way, we divided DNA sequence of Flavodoxin reductase in *Escherichia coli* into seven regions, which demonstrated a correspondence to the structural domains of the coded protein classified based on "Class," derived from the secondary structure, "Architecture," derived from the gross orientation of the secondary structures, and "Topology" defined by CATH [1].

Scherer S. et al. suggested that when the DNA sequences were analyzed by means of matrices produced by seven-order Markov Models and fragments of DNA sequence, the DNA sequences are found to contain regions with different patterns of sequence organization [2]. They supposed that there might be some as yet unknown selection pressure acting on the chromosomes and unrelated transmission of DNA information from an unrelated organism by means of the incorporation of sequences from an "alien" organelle into the nuclear genome.

In this paper, we discuss the possible incorporation of DNA sequences from other species for each region of Flavodoxin reductase, based on the homology to genes from other species.

2. Materials and Methods

2.1 materials

The DNA sequence and amino acid sequence of Flavodoxin reductase in *Escherichia coli* were downloaded from the Colibri Database [3]. The 3D protein structure was obtained from the Protein Data Bank [4].

2.2 methods

Probability of the DNA sequence of Flavodoxin reductase was analyzed by the method described in the previous paper [5] and the DNA sequence of Flavodoxin reductase was divided into the regions.

Médigue C. et al. reported that genes of *Escherichia coli* were classified into three classes (Class I, Class II, and Class III) by Fractional Correspondence Analysis and by the dynamic clustering method [6]. Genes in Class III, wherein codon usage bias was the smallest among the three classes, were the following: genes coding 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 genes coding for lambdoid phage lysogeny control protein. Also they strongly suggested that these Class III genes were mostly consisted of genes inherited by horizontal transfer [6]. Borodovsky M. et al. analyzed the DNA sequence of *Escherichia coli* with Markov models and found that only the results to be identified as the Class III genes in the genome DNA sequence, using a matrix produced with Class III genes, were allowed with acceptable accuracy. Accordingly, to explore the DNA sequences from other species for structural information, we used the third-order matrix, produced with Class III genes by Borodovsky M. et al. (*Escherichia coli* horizontally transferred genes) for our analysis [7][8]. By this methodology, when the class of a genomic DNA region is high, then the probability of the region is high. This means that analyzed DNA sequence shows similarity to the genes that were used for making the matrix. When the class of the region is low, then the probability of the region is low, meaning that this DNA sequence shows low similarity to the genes that were used for making the matrix.

For each of seven divided regions of the DNA sequence of Flavodoxin reductase, FASTA homology searches were done within the DDBJ ALL database (DDBJ release + updates Non-redundant nucleotide sequences Database) [9]. In the case where two adjacent regions are of the same class, the two regions are considered as a single region because they are assumed to be from the same species. With respect to the homology results, to prevent analysis being done on the wrong sequences, or on DNA sequences that are too short, genes were chosen whose E value (probability of the wrong sequences existing in the result) is less than 1.0 and whose ratio of overlapped length between the query sequence and the subject sequence is more than 0.5. The genes from Eukarya and genes, for which the DNA sequence is not a full-length cDNA but a fragment, were excluded. Also the genes homologous to the DNA sequences of both the FAD domain and the NADP domain were excluded since there are not meaningful for a discussion of evolution.

3. Results

3.1 DNA sequence, amino acid sequence, and 3D structure of Flavodoxin reductase in Escherichia coli

Fig.1 indicates the DNA sequence, amino acid sequence, secondary structures and the divided regions of Flavodoxin reductase in *Escherichia coli* [10][11][12]. Fig. 2 shows that the plot of class versus the position of the DNA sequence of Flavodoxin reductase as calculated with Markov models by the GeneMark program [13]. Fig. 3 represents the 3D structure of Flavodoxin reductase (PDB code 1fdr).

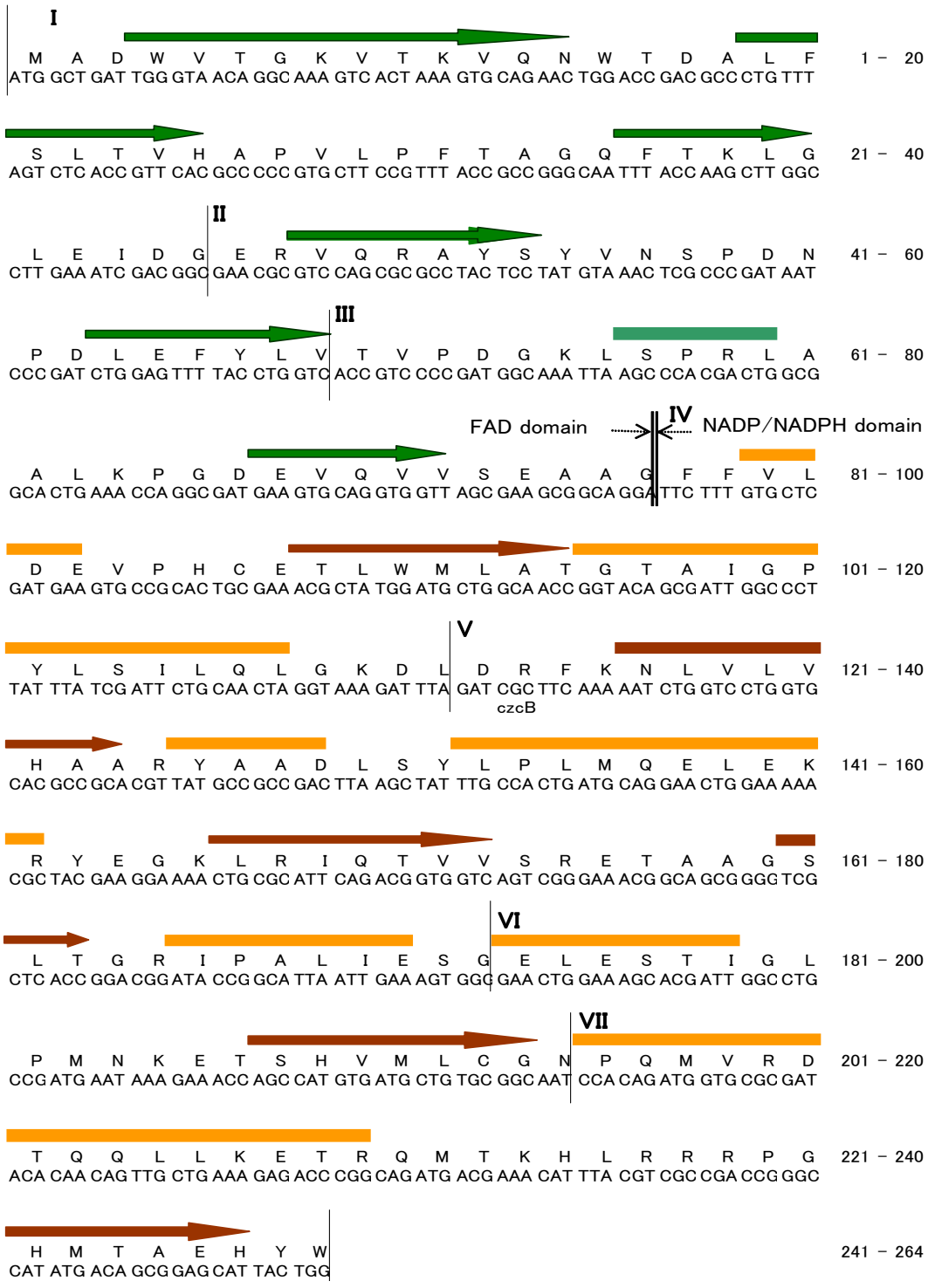


Figure 1. DNA sequence, amino acid sequence, secondary structures and the divided regions of Flavodoxin reductase in *Escherichia coli*

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 VII identify the regions divided by class.

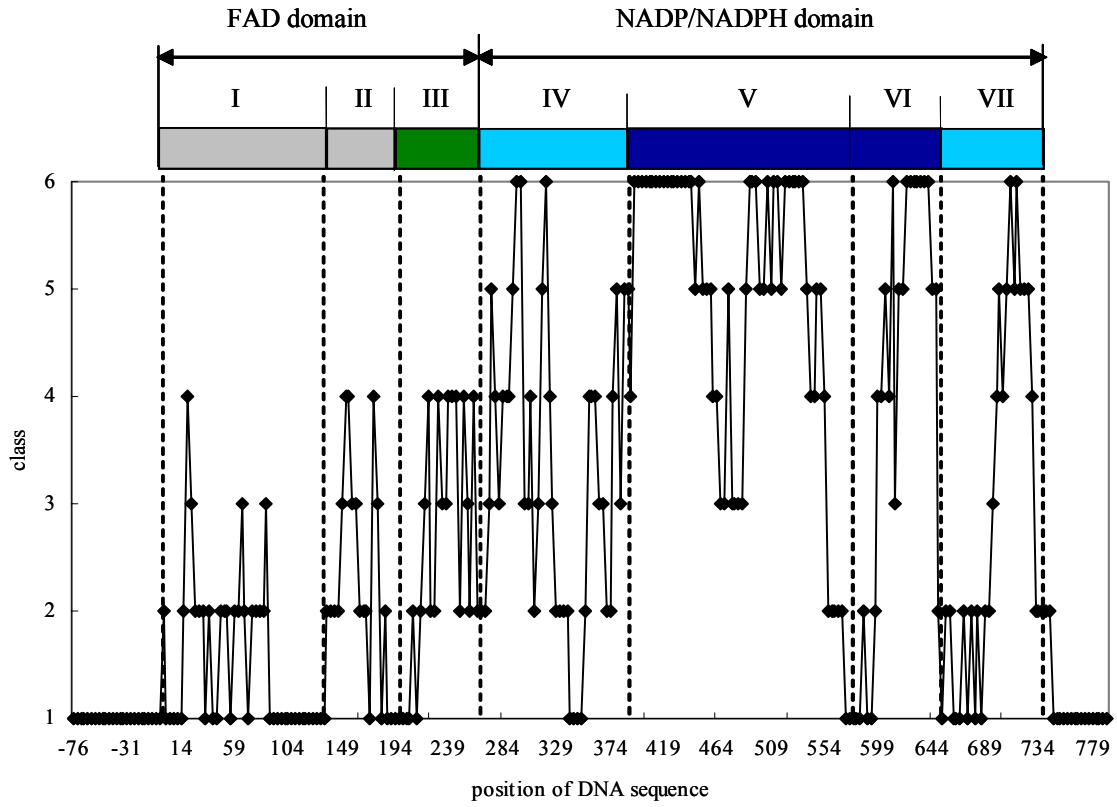
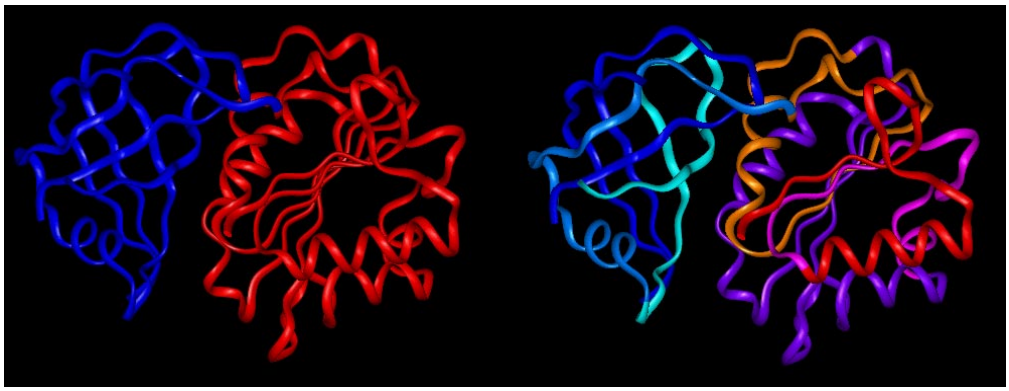


Figure 2. Regions of Flavodoxin reductase divided by probability

The regions are divided into six classes of probability as calculated by the GeneMark program. Light gray, dark green, light blue and dark blue boxes indicate the classes of regions I+II, IV, V, and VI.



(PDB code 1fdr)

Figure 3. 3D structure of Flavodoxin reductase

The figure on the left represents the domains classified by CATH. Dark blue and red ribbons show FAD and NADP domains, respectively. The figure on the right shows the seven regions divided by probability. Dark blue, light blue, sky blue, orange, violet, pink, and red represent regions, I, II, III, IV, V, VI and VII, respectively.

3.2 Homology search of the DNA sequence of Flavodoxin reductase in *Escherichia coli*

3.2.1 Regions I and II

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

The DNA sequence of region I+II was homologous to those at the N-terminal region of NADPH:ferredoxin reductase of *Xanthomonas campestris* pv. Campestris [14], *Azotobacter vinelanndii* [15], *Pseudomonas aeruginosa* PA01 [16], *Xylella fastidiosa* 9a5c [17], of the soluble methane monooxygenase gene of *Methylomonas* sp. KSWIII [18], and of protein-glutamate methylesterase of *Ralstonia solanacearum* GMI1000 megaplasmid [19] (Table 1). Genes except for those from *Azotobacter vinelanndii* and *Methylomonas* sp. KSWIII were only predicted on the basis of their DNA sequences. NADPH:ferredoxin reductase from *Azotobacter vinelanndii*, however, was cloned and expressed as a protein and its function was identified [15]. In addition, it is mentioned that the amino acid sequence at the N-terminal region was very similar to that of *Escherichia coli* [15]. The 3D structure of NADPH:ferredoxin reductase in *Azotobacter vinelanndii* was determined by X-ray crystallography [20]. When the aligned C_α atoms of the homologous amino acid residues between NADPH:ferredoxin reductase in *Azotobacter vinelanndii* and Flavodoxin reductase of *Escherichia coli* were compared, the RMSD was 1.49 Å[20]. This means that both overall structures are very similar. The soluble methane monooxygenase gene from *Methylomonas* sp. KSWIII was also cloned, expressed as a protein and its function was identified [18]. But, since NADPH:ferredoxin reductase from *Azotobacter vinelanndii* demonstrated 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, which binds FAD, would have been derived from *Azotobacter vinelanndii*.

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 methyesterase	0.31	62.19%

Table 1. Result of FASTA search, using region I+II of the FAD domain

3.2.2 Region III

No gene homologous to the DNA sequence of region III, the E-value of which was less than 1.0, was found.

3.2.3 Region IV

In the region IV, the DNA sequence shows homology to those of *lyt*B (involved in polymyxin B resistance) of *Burkholderia pseudomallei* [21], the conjugal transfer gene E (*tra*E) of *Escherichia coli* pKM101 [22] or IncN plasmid R46 [23], the hypothetical protein MmyD of *Streptomyces coelicolor* plasmid SCP1 [24] and an unidentified sequence of *Hydrogenophaga pseudoflava* [25] or *Sphingomonas paucimobilis* [26] (Table 2). Comparing the restriction enzyme maps between *Escherichia coli* pKM101 and IncN plasmid R46, *Escherichia coli* pKM101 was noted to be derived by deletetion of the regions related to heavy metal and drug resistance from the IncN plasmid R46[27]. Genes coded on plasmids can be transferred via cell-to-cell conjugation from one species to another. Conjugal transfer gene E (*tra*E) is a gene 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 would be coming from horizontally transferred genes, which were used for the model made by Borodovsky M. et al.. Therefore, the DNA sequence of region IV, which binds NADP/NADPH, would have been derived from the *Escherichia coli* plasmid pKM101.

GI	species	gene	E value	identity
AF098521	*Burkholderia pseudomallei*	*lyt*B (involved in polymyxin B resistance)	0.21	77.78%
AY046276	IncN plasmid R46	*tra*E (conjugation)	0.32	70.00%
U09868	*Escherichia coli* pKM101	*tra*E (conjugation)	0.46	70.00%
AL590464	*Streptomyces coelicolor* plasmid SCP1	hypothetical protein, MmyD	0.67	70.59%
U55037	*Hydrogenophaga pseudoflava*	unidentified sequence	0.71	64.29%
AB033664	*Sphingomonas paucimobilis*	unidentified sequence	0.86	62.96%

Table 2. Result of FASTA search, using region IV of the NADP/NADPH domain

3.2.4 Regions V and VI

Since regions V and VI show the same CLASS, the two regions were regarded as a single region. Therefore, the homology search was done for region V+VI.

The DNA sequence of region V+VI was homologous to those of *czc*B (membrane fusion protein) of *Ralstonia* sp. CH34 plasmid pMOL30 [28] and *Alcaligenes* sp. [29] (Table 3). Since the class of this region is high, it is considered likely that the DNA sequence of this region would have been come from horizontally transferred genes, which were used for the matrix produced by Borodovsky M. et al. Consequently, the DNA sequence of the region V+VI, which binds NADP, would have been derived from the *Ralstonia* sp. CH34 plasmid pMOL30. The gene from the *Ralstonia* sp. CH34 plasmid pMOL30 that showed homology to the DNA sequence of this region is the *czc*B gene, which may function as a cation-binding subunit and act as a funnel in the mechanism of the cation-proton antipoter [28].

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%

Table 3. Result of FASTA search, using region V+VI of the NADP/NADPH domain

3.2.5 Region VII

The DNA sequence of region VII was homologous to those of DNA repair exonuclease (*sbc*D) of *Bacillus subtilis* [30], *rhc*U of *Bradyrhizobium japonicum* [31] and orf17 of *Streptococcus pneumoniae* bacteriophage Cp-1 [32] (Table 4). These genes, except for those from *Streptococcus pneumoniae* bacteriophage Cp-1, were only predicted based on the DNA sequences. The orf17 gene homologous to the DNA sequence of this region was cloned, expressed as a protein and deduced to be the tail protein by the similarity of its N-terminal amino acid sequence [32]. Since the class of this region is high, it is considered likely that the DNA sequence of this region would have been come from horizontally transferred genes, which were used for the matrix produced by Borodovsky M. et al. Therefore, the DNA sequence of region VII, which binds NADP, would have been derived from the *Streptococcus pneumoniae* bacteriophage Cp-1.

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	*Bacteriophage* Cp-1	orf17	0.32	62.00%

Table 4. Result of FASTA search, using region VIII of the NADP/NADPH domain

4. Discussion

It was demonstrated that the DNA sequence of Flavodoxin reductase in *Escherichia coli* was divided into seven regions as shown in Fig. 2 [5], and that the divided regions also showed a correspondence to the domains, which corresponded to those defined by CATH, of the 3D structure of Flavodoxin reductase determined by X-ray crystallography. Moreover, they also corresponded to the gross secondary structures of the protein, as shown Fig. 3.

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. The three genes, which the DNA sequence of the NADP/NADPH domain is elucidated to have been derived from, have the following characteristics; *Ralstonia* sp. CH34 plasmid pMOL30 *czc*B (membrane fusion protein) is one among 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-terminal regions and these mutants could be expressed at high levels in the host cells and function properly [33]. This suggests that the DNA sequence of phage tail protein and DNA fragments from other species might be fused together easily when they become incorporated into a cell. Thus, the regions I+II, IV, V+VI, and VII 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 plasmid pMOL30 and orf17 from *Streptococcus pneumoniae* bacteriophage Cp-1, respectively.

We 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 encoded on plasmids or bacteriophages. 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 the NADPH:ferredoxin reductase from *Azotobacter vinelanndii* might have occurred to become the FAD domain. The DNA sequences might have been incorporated to become the NADP domain from the genes related to heavy metal resistance and to conjugation encoded on a plasmid, and that, of function related to tail protein, encoded on a bacteriophage (horizontally transfer). 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. Therefore, 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 correspondence with the domains of the 3D structure.

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 our analysis. On the level of amino acid sequences, it is mentioned that the amino acid sequence at the N-terminus region of NADPH:ferredoxin from *Azotobacter vinelanndii* was very similar to that of Flavodoxin reductase from *Escherichia coli* [15]. The 3D structures of Flavodoxin reductase from *Escherichia coli* and NADPH:ferredoxin reductase from *Azotobacter vinelanndii* were determined by X-ray crystallography. And the RMS deviation (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 1.49 Å [20]. This means that the overall structure between them is very similar. These results also suggested that there exists a relationship among the 3D structure, amino acid sequence, and DNA sequence corresponding to the domains of the 3D structure, between Flavodoxin reductase from *Escherichia coli* and NADPH:ferredoxin reductase from *Azotobacter vinelanndii*. Taken together, Flavodoxin reductase from *Escherichia coli* and NADPH:ferredoxin reductase from *Azotobacter vinelanndii* would have been incorporated into their respective genomes and then the 3D structure or domains of the 3D structure have not been changed in evolution. As a result, differences between the DNA sequences of the gene from *Escherichia coli* and that from *Azotobacter vinelanndii* might have been increased.

It is elucidated 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 reductase *from Azotobacter vinelanndii*, but the 3D structures of proteins from both species are conserved. By the analysis of the G+C content at the third codon position of synonymous codon,

we suggested that there might be the genes in *Pyrococcus horikoshii* OT3 in which the DNA sequences would been changed but the 3D protein structures would not been affected in evolution [34]. 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 3D structure or the domains of the 3D structure and DNA sequence of a gene, and this represents the information accumulated in evolution, which had been encoded on the DNA sequence.

References

[1] C.A. Orengo et al., *Structure*, **5**, 1093-1108 (1997).
[2] S. Scherer et al., *Proc. Natl. Acad. Sci. USA.*, **95**, 10026-10031 (1994).
[3] http://www.rcsb.org/pdb/
[4] http://genolist.pasteur.fr/Colibri/
[5] Y. Ohfuku et al., *CBIJ.,* **2**, 119-136 (2002).
[6] C. Médigue et al., *J. Mol. Biol.*, **222**, 851-856 (1991).
[7] M. Borodovsky et al., *Nucleic Acids Res.*, **23**, 3554-3562 (1995).
[8] http://opal.biology.gatech.edu/GeneMark/downloads.html
[9] http://spiral.genes.nig.ac.jp/homology/
[10] V. Bianchi et al., *J. Bacteriol.*, **175**, 1590-1595 (1993).
[11] M. Ingelman et al., *J. Mol. Biol.*, **268**, 147-157 (1997).
[12] Y. Ohfuku et al., *CBIJ.,* **2**, 119-136 (2002).
[13] M. Borodovsky and J. McIninch, *Computers Chem.*, **17,** 123-133 (1993).
[14] A. C. Da Silva et al., *Nature*, **417**, 459-463 (2002).
[15] J. M. Isas et al., *J. Biol. Chem.*, **270**, 21258-21263 (1995).
[16] C. K. Stover et al., *Nature*, **406**, 959-964 (2000).
[17] A. J. Simpson et al., *Nature*, **406**, 151-157 (2000).
[18] T. Shigematsu et al., *Appl. Environ. Microbiol.*, **65**, 5198-5206 (1999).
[19] http://sequence.toulouse.inra.fr/R.solanacearum.html
[20] G. Prasad et al., *Protein Sci.*, **7**, 2541-2549 (1998).
[21] M. N. Burtnick and D. E. Woods, *Antimicrob. Agents Chemother.*, **43**, 2648-2656 (1999).
[22] S. C. Winans and G. C. Walker, *J. Bacteriol.*, **161**, 402-410 (1985).
[23] A. M. Brown and N. S. Willetts, *Plasmid*, **5**, 188-201 (1981).
[24] M. Redenbach et al., *J. Bacteriol.*, **180**, 2796-2799 (1998).
[25] Y. M. Kim and S. N. Lee, direct Submission to NCBI.
[26] X. Peng et al., *Appl. Environ. Microbiol.*, **64**, 2520-2527 (1998).
[27] P. J. Langer and G.C. Walker, *Mol. Gen. Genet.*, **182**, 268-272 (1981).
[28] C. Rensing et al., J. Bacteriol., 179, 6871-6879 (1997).
[29] T. Kunito et al., *Biosci. Biotechnol. Biochem.*, **60**, 699-704 (1996).
[30] N. Medina et al., *Microbiology*, **143**, 3305-3308 (1997).
[31] M. Gottfert et al., *J. Bacteriol.*, **183**, 1405-1412 (2001).
[32] A. C. Martin et al., *J. Virol.*, **70**, 3678-3687 (1996).
[33] X. Carbonell and A. Villaverde, Biochem. Biophys. Res. Commun., 244, 428-433 (1998).
[34] Y. Ohfuku et al., *CBIJ.,* **2,** 58-73 (2002).