Protein Evolution
Based on Analysis of DNA Sequences

Division of Agriculture and Forestry
Doctoral Degree Program in Agricultural Sciences
University of Tsukuba

Yuko Ohfuku
Contents

1 Introduction 1

2 Materials and Methods 7
   2.1 Definition of orthologous genes 7
   2.2 The G+C content and the G+C content at the third codon position 7
   2.3 Division of a DNA sequence by using a Markov model 9
   2.4 Definitions of class, region and CLASS 10

3 Lateral genes in the genomic DNA sequence 13
   3.1 The correlation between the G+C content of the genomic DNA sequences and the G+C content at the third codon position (GC3 content) for all ORFs of the genomic DNA sequences 13
   3.2 The correlation between the G+C content of the genomic DNA sequences and the G+C content at the third position of the three types of codons found in orthologous genes 16
   3.3 Distribution of the G+C content at the third codon position of synonymous codons in the orthologous genes of Pyrococcus horikoshii OT3 19
   3.4 Characteristics of the high and low GC3 content regions 22

4 Relationship between DNA sequence and three-dimensional structure 29
   4.1 Analysis of probability of DNA sequence 29
      4.1.1 Flavodoxin reductase 29
      4.1.2 Heat shock protein (grpE) 32
      4.1.3 Flavin oxidoreductase 32
      4.1.4 Integrase/recombinase zerD 37
      4.1.5 Endonuclease III 37
      4.1.6 Elongation factor Tu (tufB) 42
4.2 Comparison between the divided regions and the domain classified in accordance with the definition of CATH 42

5 Structural evolution of protein 47
5.1 Flavodoxin reductase of *Escherichia coli* 47
  5.1.1 DNA sequence, amino acid sequence, three-dimensional structure and the divided regions 47
  5.1.2 Homology search with the DNA sequences of the divided regions 47
5.2 Heat shock protein (*grpE*) of *Escherichia coli* 54
  5.2.1 DNA sequence, amino acid sequence, three-dimensional structure and the divided regions 54
  5.2.2 Homology search with the DNA sequences of the divided regions 54
  5.2.3 Duplication of gene 59

6 Discussion 65

Acknowledgements 75

References 77

Appendix 91