Gene Has Its Inherent Significantly Repetitive Tuples

Nobuyuki Uchikoga uchiko@genji.c.u-tokyo.ac.jp

Akira Suyama suyama@dna.c.u-tokyo.ac.jp

Department of Life Sciences, The University of Tokyo
3-8-1 Komaba, Meguro-ku, Tokyo 153, Japan

1 Introduction

Nonenzymatic synthesis of nucleic acid yields only short oligomer compared with enzymatic synthesis using the modern machinery. It leads us to a question how nucleic acid chains, which are long enough to encode the functional polypeptide chains, were created in primordial soup. To answer this question, S. Ohno proposed a model of primordial sequences composed of repeats of short base oligomers [1]. This model is very attractive. However, the universality of the model still remains open to question because he derived the model from base sequences of some genes with unusually regular oligomeric repeats.

In 1996, we examined the universality of the model by analyzing base sequences of about eighty genes with different functions classified into the major three biospheres (Archaea, Bacteria, Eukaryotae) on the phylogenetic tree of life. This examination showed that the model of the primordial coding sequences is likely to be universal because homologous tuples appear frequently in both coding and noncoding regions of the present living organisms [2]. However, this conclusion did not indicate that specific base sequences of homologous tuples exist universally in both coding and noncoding regions of modern genes. In our previous work, we then investigated the distribution of sets of significantly repetitive tuples with the same base sequence. The conclusion of the previous work showed that significantly repetitive tuple with the same base sequence is specific to each gene and each biosphere [3].

Although the conclusion of the previous work, significantly repetitive tuples with homologous base sequence in both coding and noncoding regions can be regarded as vestiges of primordial genes because base sequences of genes could have been changed while primordial sequences have diverged to modern genes. These conclusions above led us to examine the distribution of significantly repetitive homologous tuples by analyzing base sequences of genes classified into three biospheres.

2 Method

For each tuple, the significance, x, of the frequency of the occurrence of a tuple in a sample base sequence is given by

$$x = \frac{f_0 - Np}{\sqrt{Np(1-p)}},$$

where f_0 is the observed number of a tuple, N is the total number of a tuple in a base sequence examined, and p is the probability of the occurrence of a tuple in random sequences whose base composition is same with a sample sequence. We defined the significantly repetitive tuple (SRT) as a tuple with $x > 3$.

Table 1: The ratio of the number of the homologous whole gene SRTs shared with ALL genes (%) to all SRTs existing in each biosphere. (Boldface is the ratio larger than 45).

Genes chosen from data base randomly

	5-tuple	6-tuple	7-tuple	The number of genes
Archaea	11.8	20.3	0.3	21
Bacteria	0.1	12.0	0.7	44
Eukaryotae	6.7	10.3	1.0	21

Nine contiguous genes on *E. coli* genome around dnaA
(f135, gryB, recF, dnaN, dnaA, rmpH, rnpA, yidC, thdF)

5-tuple	6-tuple	7-tuple	The number of genes
1.0	14.4	0.7	9

Heat Shock Protein 70 Family

	5-tuple	6-tuple	7-tuple	The number of genes
Archaea	**94.3**	**79.9**	28.1	3
Bacteria	**55.8**	**80.6**	**55.6**	8
Eukaryotae	17.5	**47.4**	18.4	21

3 Results and Discussion

We analyzed three sets of base sequences of gene. One is the set of about eighty genes chosen from database randomly. Another is the set of nine contiguous genes around dnaA on *E.coli* genome. The last is Heat Shock Protein 70 family in each biosphere.

We paid special attention to SRTs observed in both coding and noncoding regions of each gene because the primordial genes could become not only coding sequences but also noncoding sequences of modern genomes. We also paid attention to whole gene SRTs shared with all genes because primordial sequences might have diverged to modern genes. It is believed that base sequences of SRTs could have been changed while primordial sequences diverging. In this work, we define homologous tuples are sets of tuples with base mismatches less than two bases. These whole gene SRTs can be regarded as vestiges of primordial genes. Table 1 shows the distribution of whole gene SRTs shared with all genes.

Table shows that more homologous whole gene SRTs are shared with Heat Shock Protein 70 family than two other gene sets. These homologous whole gene SRTs are in not only coding region but also noncoding region. Such SRTs appear frequently in one gene family. These facts imply that the homologous whole gene SRTs can be regarded as vestiges of the primordial words. The gene family may have a specific set of primordial genes whose base sequences are similar to each other.

Acknowledgements

This work was supported in part by a Grant-in-Aid for Scientific Research on Priority Areas from The Ministry of Education, Science, Sports and Culture in Japan.

References

[1] Ohno, S., Repeats of base oligomers as the primordial coding sequences of the primeval Earth and their vestiges in modern genes, *J. Mol. Evol.*, 20:313–321, 1984.

[2] Uchikoga, N and Suyama, A., Vestiges of primordial word in base sequences of modern genomes, *Genome Informatics 1996*, Universal Academy Press, 242–243, 1996.

[3] Uchikoga, N and Suyama, A., Distribution of Significantly Repetitive Tuples Implying Primordial genes, *Genome Informatics 1997*, Universal Academy Press, 350–351, 1997.