Distribution of Significantly Repetitive Tuples Implying

Nobuyuki Uchikoga **Akira Suyama**
uchiko@genji.c.u-tokyo.ac.jp suyama@dna.c.u-tokyo.ac.jp
Department of Life Sciences, The University of Tokyo
3-8-1 Komaba, Meguro-ku Tokyo 153, Japan

Abstract

We studied the base sequences of tuples frequently occurring in both coding and noncoding regions of several modern genomes classified into the three major biospheres of the phylogenetic tree of life. These significantly repetitive tuples were found to be specific not only to genes but also to the biospheres.

1 Introduction

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

We thus examined the universality of this model by analyzing base sequences of several genomes classified into the major three biospheres (Archaea, Bacteria, Eucarya) on the phylogenetic tree of life. The 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]. This conclusion has led us to examine actual base sequences of the homologous tuples occurring with significantly high frequency.

2 Method

For each tuple, the significance, x, of the frequency of the occurrence of a tuple 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. We defined the significantly repetitive tuple (SRT) as a tuple with $x > 3$.

3 Results and Discussion

We paid special attention to SRTs observed in both coding and noncoding regions. These whole gene SRTs are regarded as vestiges of the primordial genes because the primordial genes could evolve into not only coding regions but also noncoding regions of modern genomes. Tables 1 and 2 show the distribution of the whole gene SRT.

Table 1. The ratio of the number of the whole gene SRTs shared with n genes (%)

biosphere	$n = 1$	$n = 2$	$n \geq 3$
Archaea	98	2	0
Bacteria	85	15	0
Eucarya	95	5	0

Table 2. The ratio of the number of the whole gene SRTs shared with n biospheres (%)

biosphere	$n = 1$	$n = 2$			$n = 3$
		Archaea/Bacteria	Archaea/Eucarya	Bacteria/Eucarya	
Archaea	86	5	7		2
Bacteria	90	3		5	2
Eucarya	83		7	7	2

Table 1 shows that the whole gene SRT is specific to each gene. Very few whole gene SRTs are shared with genes. In addition, Table 2 shows that the whole gene SRT is specific to each biosphere. These facts imply that each gene of the modern genome may evolve from a different primordial gene through some mechanisms preserving vestiges of the primordial coding sequences.

Acknowledgments

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] S. Ohno, “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] N. Uchikoga and A. Suyama, “Vestiges of Primordial Words in Base Sequences of Modern Genomes,” *Proceedings of Genome Informatics Workshop VII*, pp. 242–243, 1996.