Distances between Completely Matched Tuples in the Modern Gene Sequences

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 153-8902, Japan

1 Introduction

Modern genome sequences are the products as a consequence of gene evolution. When we consider the gene evolution, it is necessary to investigate a universal feature of base sequences of the modern genes. We have investigated the characteristics of base sequences of modern genes using statistical analysis. In our previous works, we revealed that the base sequences of the modern genes universally have statistically significant repetitive short tuples [1]. These significant repetitive tuples (SRTs) exist not only in coding region but also in noncoding region and the lengths of SRTs are not multiple of the codon length. Each gene has a specific set of the base sequences of SRTs in the whole gene consisting of the coding and noncoding regions [2]. A set of genes encoding proteins with different functions contains few common SRTs. On the other hand, a gene family has a number of common SRTs.

It is possible to say that the base sequences of SRTs are influenced by the codons used frequently in the coding region. If the base sequences of the repetitive short tuples are related to the sequences of codons, our results reflect the preference of codon usage. We have not yet studied the actual location of SRTs in the reading frame. Instead, in this work, we analysed distances of completely matched tuples to investigate whether or not the location of tuples with the same base sequence in coding region is related to the reading frame.

2 Method

The distances between the tuples with the identical sequences, D, are calculated for all possible tuples, and they are divided by k ($k = 3, 4, 5, 6$). The normalized distribution of m, where $m = D \bmod k$, is given by

$$d(m) = \frac{N(m) - N_{random}(m)}{\sqrt{\sigma_{random}(m)}},$$

where $N(m)$ is the number of the pairs with m found in a sample gene, and $N_{random}(m)$ is that observed in randomly generated sequences with the standard deviation $\sqrt{\sigma_{random}(m)}$. We generated 1000 random sequences with the same base composition and the same length as the corresponding gene. For example, when distances of each pair are divided by three, $N(0)$ is the observed number of pairs whose distances of such pairs are multiple of three. Then $d(0)$ indicates the statistical significance of the occurrence of the pairs whose distances are multiple of three. In this work, we have analyzed tuples with five base length. The sequences having remarkable repetitions are excluded.

3 Results

Fig. 1 shows the normalized distributions of m. We recognized the statistically significance of the tuple pairs with the multiples of 3 intervals (See Fig. 1(a) for $m = 0$ and Fig. 1(d) for $m = 0$, 3).

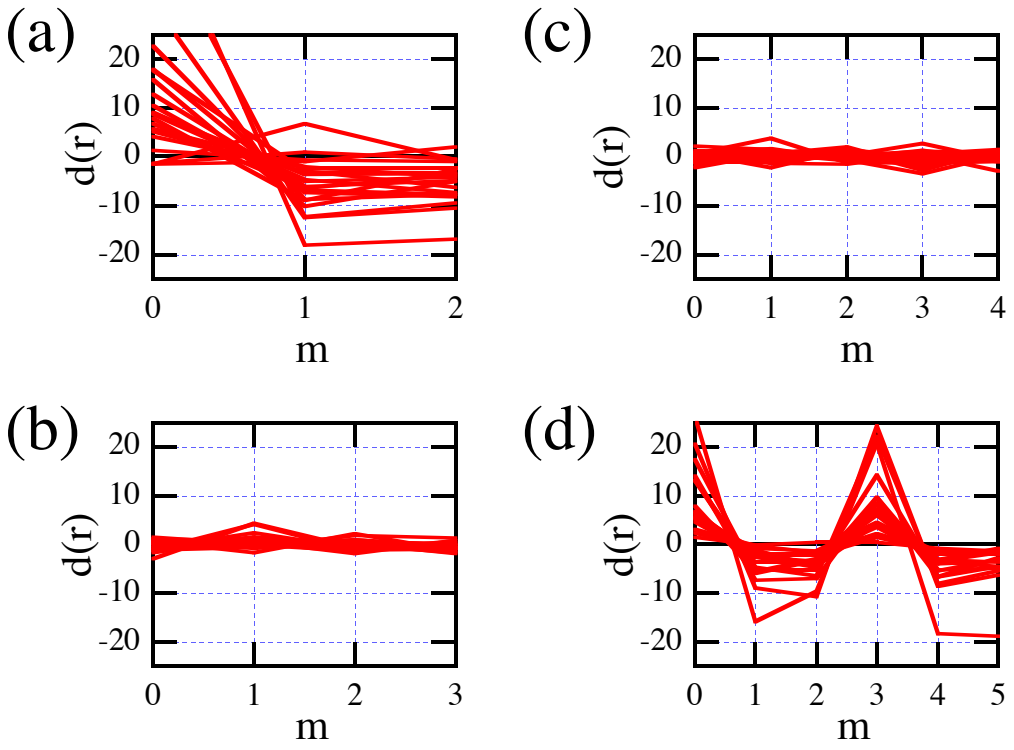

Figure 1: Normalized distribution of m, $d(m)$, are shown. Twenty one Archaea coding regions are analyzed as sample data. (a): $k = 3$, (b): $k = 4$, (c): $k = 5$, (d): $k = 6$. (See text for m, k and $d(m)$).

Fig. 1(b) and (c) do not show any discernible tendency. Although we here deal only with Archaea coding regions, coding regions of other biospheres (Bacteria and Eukarya) also show a similar profile (data not shown).

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] Uchikoga, N. and Suyama, A., Vestiges of primordial words in base sequences of modern genomes," *Genome Informatics 1996*, 242–243, 1996.

[2] Uchikoga, N. and Suyama, A., Gene has its inherent significantly repetitive tuples, *Genome Informatics 1998*, 388–389, 1998.