

Significant Deviation in the Configuration of Tandem Repeats in Prokaryotic Genomes

Shintaro Hirayama

gs06415@eit.hirosaki-u.ac.jp

Satoshi Mizuta

slmizu@cc.hirosaki-u.ac.jp

Faculty of Science and Technology, Hirosaki University, Hirosaki ,Aomori 036-8561, Japan

Keywords: tandem repeat, whole-genome duplication, prokaryotic genome, Kullback-Leibler divergence

1 Introduction

Various studies of whole-genome duplication (WGD) have been done on eukaryotic genomes, and WGD as well as gene duplication is being thought to be one of the important processes in genome evolution for some eukaryotic lineages such as yeast, fishes, and vertebrates [4]. As for prokaryotic species, however, only a few studies on *Escherichia coli* [2, 5] and *Anabaena* [3] have been reported. Furthermore, the former two studies had been performed before the genome was completely sequenced.

In this study, we seek the evidences of WGD on completely sequenced prokaryotic genomes based on methods of bioinformatics and statistical analysis.

2 Materials and Method

Genome sequences were obtained from GenBank[6] and tandem repeats (TRs) were detected by tandem repeats finder (TRF)[1]. All-against-all pairwise alignment was performed on the detected TRs using SSEARCH[7] with the default parameter settings and a threshold E -value = 10^{-3} . A pair of TRs which have the overlap of more than 50% to the longer sequence was defined as an equivalent TR-pair.

The central angles on the circular genomes of all the equivalent TR-pairs were calculated and their distributions were measured at a certain class interval. The discrepancy between the observed distribution, $P =$

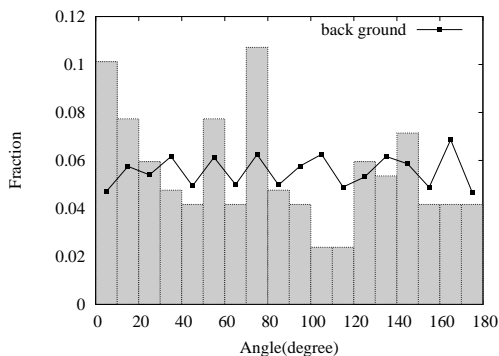


Figure 1: Central angle distribution of equivalent TR-pairs with back ground

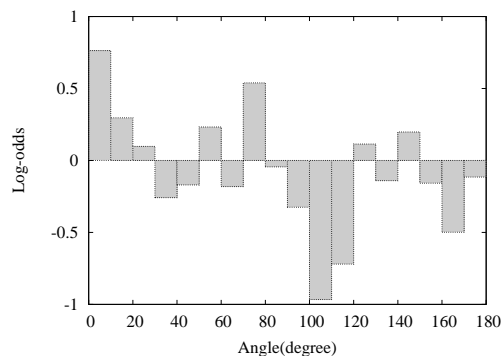


Figure 2: Log-odds ratio between the observation and the back ground

(p_1, p_2, \dots, p_k) , and that of the back ground, $Q = (q_1, q_2, \dots, q_k)$, was evaluated by the Kullback-Leibler (KL) divergence defined by

$$D(P \parallel Q) = \sum_{i=1}^k p_i \log \frac{p_i}{q_i}, \quad (1)$$

where the back ground distribution was determined by the averaged angle distribution of randomly arranged 10,000 sets of TR-pairs preserving the number of TRs, the equivalence relations between TRs, and the loci of the TRs on the genomes.

The statistical significance of the discrepancy of the observed distribution from the back ground one is certified by a p -value estimated to be $\sim n/10^4$, where n is the number of the random sets which have the divergence greater than the observed one.

3 Results and Discussion

Figure 1 shows the distribution of the central angle of the equivalent TR-pairs detected on *E. coli* K12 genome. If the TRs were distributed completely at random on the genome, a flat distribution would be observed. Although the observed distribution is far from flat, it does not necessarily indicate something noteworthy, because the loci of TRs are not uniform on the genome for some reasons independent of WGD. Consequently, we took the log-odds ratio defined by $\log(p_i/q_i)$ into consideration. Figure 2 depicts the angle distribution of the log-odds ratio. We can recognize apparent discrepancy between the observed and the back ground distributions as well as some periodicity, which suggests the existing of WGD on the genome.

Figure 3 shows the distribution of KL divergence of the random sets with the same back ground distribution as before. The observed divergence is 0.0802 and the p -value is estimated to be 0.0338, which shows that the departure from the null hypothesis is assessed at the 5% significance level. One of the reasonable explanation of the deviation is WGD.

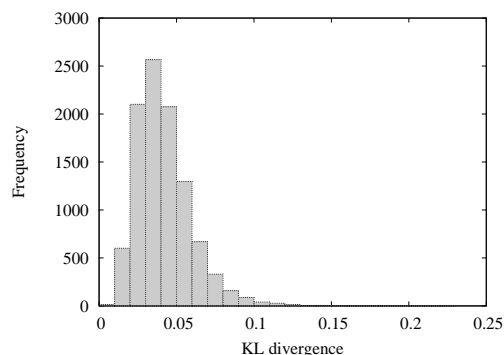


Figure 3: Distribution of KL divergence of random sets

References

- [1] Benson, G., Tandem repeats finder: A program to analyze DNA sequences, *Nucleic Acids Res.*, 27:573–580, 1999.
- [2] Kunisawa, T. and Otsuka, J., Periodic distribution of homologous genes or gene segments on the *Escherichia coli* K12 genome, *Protein Seq. Data Anal.*, 1:263-267, 1988.
- [3] Sugaya, N., Sato, M., Murakami, H., Imaizumi, A., Aburatani, S., and Horimoto, K., Causes for the large genome size in a *Cyanobacterium Anabaena* sp. PCC7120, *Genome Informatics*, 15:229–238, 2004.
- [4] Van de Peer, Y., Computational approaches to unveiling ancient genome duplications, *Nat. Rev. Genet.*, 5:752–763, 2004.
- [5] Zipkas, D. and Riley, M., Proposal concerning mechanism of evolution of the genome of *Escherichia coli*, *Proc. Natl. Acad. Sci. USA*, 72:1354–1358, 1975.
- [6] <http://www.ncbi.nlm.nih.gov/>
- [7] <http://fasta.bioch.virginia.edu/>