Gene-Distribution Patterns on Cyanobacterial Genomes

Nobuyoshi Sugaya sugaya@ims.u-tokyo.ac.jp
Hiroo Murakami hiroo@ims.u-tokyo.ac.jp
Makihiko Satoh makihiko@ims.u-tokyo.ac.jp
Sachiyo Aburatani sachiyo@ims.u-tokyo.ac.jp
Katsuhisa Horimoto khorimot@ims.u-tokyo.ac.jp
Human Genome Center, Institute of Medical Science, University of Tokyo, Tokyo 108-8639, Japan

Keywords: genome size, gene distribution, gene-location distance, cyanobacteria

1 Introduction

In the phylum Cyanobacteria, complete genome sequences have been already reported in seven species and strains [1, 4, 5, 6, 7, 8]. Very interestingly, only one speices of these cyanobacteria, *Anabaena* sp. PCC7120 [5] (hereafter Anabaena), has considerably large genome size of approximately 6.4 Mb, while the remaining species have the genome sizes of 1.7-3.6 Mb to be in the range only from about a quarter to a half of the size of the Anabaena genome. Genome size difference in bacterial species has been well known in the relation between parasitic species such as *Mycoplasma* and *Buchnera* and their close relatives. In these cases, genome-size difference has been ascribed to the genome-size reduction in parasitic species due to their life style of obligate parasitism. On the contrary, cyanobacterial species whose complete genomes have been sequenced so far are all free-living species. In this study, we approach the genome-size difference in these cyanobacterial species by investigating patterns of gene distribution on the genomes of three cyanobacteria, Anabaena, *Synechocystis* sp. PCC6803 (*ca.* 3.6Mb) [1] (hereafter Synechocystis) and *Thermosynechococcus elongatus* BP-1 (*ca.* 2.6Mb) [6].

2 Method and Results

In order to investigate patterns of gene distribution on three cyanobacterial genomes, we adopt a metric of gene-location distance [2, 3] to the gene-configuration comparison between the Anabaena genome and the genomes of Synechocystis and *T. elongatus*. Gene-location distance is a measure of the dissimilarity of a pair of genes, relative to the configuration of other pairs of genes on two circular genomes compared [2, 3]. For the present purpose, the Anabaena genome is divided into two halves and each gene configuration on a half of the Anabaena genome is compared with that on other two genomes (Fig. 1).

Common orthologs among three species are identified with the use of the program BLASTP. The number of common orthologs thus identified is 1449, and gene-location distances are calculated with respect to these ortholog pairs between a half region of the Anabaena genome and whole region of another genome. Then, the smallest values of gene-location distance are plotted against the cutting angles on the Anabaena genome. The plots shown in Fig. 2 indicate that when cut at an angle, gene configuration on the half region becomes most similar to that on whole region of another genome. As a result, gene configuration on the region of 260°-80° in the Anabaena genome is similar to that on the Synechocystis genome, and that on the region of 80°-260° is similar to that on the *T. elongatus* genome.

3 Discussion

Anabaena is a filamentous, heterocyst-forming and nitrogen-fixing cyanobacterium, while Synechocystis and *T. elongatus* are unicellular and non-nitrogen-fixing cyanobacteria [4, 5, 6]. It is of interest that when the Anabaena genome is divided into two regions of 260°-80° and 80°-260°, a larger number of house-keeping protein genes involved in cell maintenance such as protein synthesis and photosynthesis

are encoded on the former region, on the other hand, a number of Anabaena-specific protein genes involved in heterocyst differentiation and nitrogen fixation are on the latter region. These distribution patterns of functional categories are statistically significant by chi-square test for independence ($P < 10^{-6}$). The results obtained from the plots of gene-location distance and the distribution patterns of functional categories imply that the contemporary Anabaena genome might have increased its genome size and amplified the gene repertoire by an event such as whole-genome duplication and a fusion of two bacterial genomes during its course of evolution.

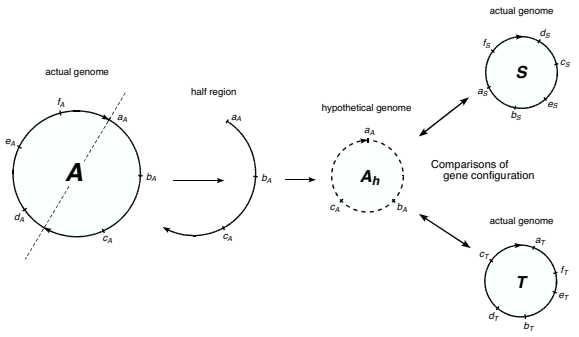

Figure 1: Illustration of the procedure for comparing gene configuration of ortholog pairs between a half region of the Anabaena genome and whole region of another genome. The actual Anabaena genome (shown by solid circle **A**) is divided into two regions by cutting it at each angle for the location of orthologs. A hypothetical genome (shown by broken circle $\mathbf{A}_h$) is constructed from one of these two regions. Gene-location distances for ortholog pairs (a, b and c in this example) are calculated between a hypothetical genome and the actual Synechocystis (shown by solid circle **S**) or *T. elongatus* (shown by solid circle **T**) genome.

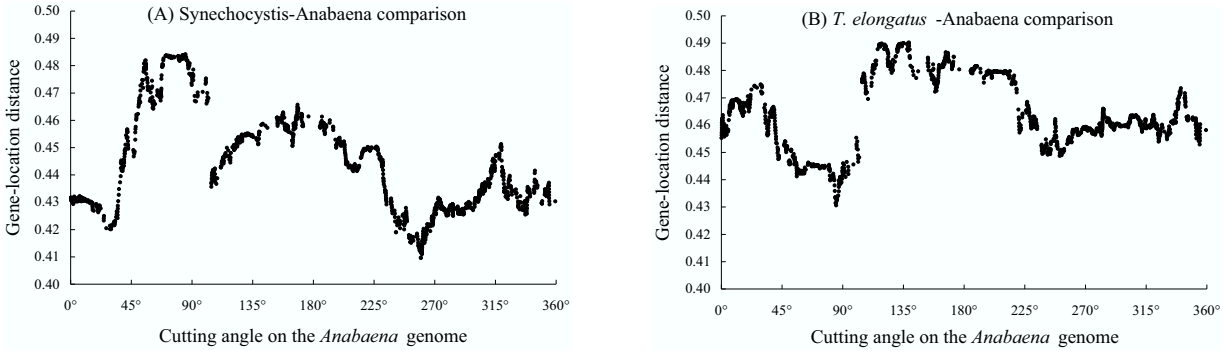

Figure 2: The plots of gene-location distance against the cutting angle on the Anabaena genome.

References

[1] Dufresne, A. *et al.*, Genome sequence of the cyanobacterium *Prochlorococcus marinus* SS120, a nearly minimal oxyphototrophic genome, *Proc. Natl. Acad. Sci. USA*, 100:10020–10025, 2003.

[2] Horimoto, K. *et al.*, A method for comparing circular genomes from gene locations: application to mitochondrial genomes, *Bioinformatics*, 14:789–802, 1998.

[3] Horimoto, K. *et al.*, Comprehensive comparison between locations of orthologous genes on archaeal and bacterial genomes, *Bioinformatics*, 17:791–802, 2001.

[4] Kaneko, T. *et al.*, Complete genomic sequence of the filamentous nitrogen-fixing cyanobacterium *Anabaena* sp. Strain PCC7120, *DNA Res.*, 8:205–213, 2001.

[5] Kaneko, T. *et al.*, Sequence analysis of the genome of the unicellular cyanobacterium *Synechocystis* sp. Strain PCC6803. II. Sequence determination of the entire genome and assignment of potential protein-codimg regions, *DNA Res.*, 3:109–136, 1996.

[6] Nakamura, Y. *et al.*, Complete genome structure of the thermophilic cyanobacterium *Thermosynechococcus elongatus* BP-1, *DNA Res.*, 9:123–130, 2002.

[7] Palenik, B. *et al.*, The genome of a motile marine *Synechococcus*, *Nature,* 424:1037–1042, 2003.

[8] Rocap, G. *et al.*, Genome divergence in two *Prochlorococcus* ecotypes reflects oceanic niche differentiation, *Nature,* 424:1042–1047, 2003.