

Distance Structure of Homologous Gene Clusters in *Cyanobacteria*

Naobumi Sasaki

naobumi@bio.c.u-tokyo.ac.jp

Naoki Sato

naokisat@bio.c.u-tokyo.ac.jp

Graduate School of Arts and Science, University of Tokyo, 3-8-1 Komaba, Meguro-ku,
Tokyo 153-0041, Japan

Keywords: genome structure, comparative genomics, *cyanobacteria*, gene cluster

1 Introduction

Rearrangement of genomes reflects evolutionary history of the genomes, which have left conspicuous traces on the genomes that are detected as conserved adjacency of homologous genes within different genomes. Such conserved ordering of genes are recognized as synteny. To find such conserved genome structure, various methods of comparative genomics are used extensively.

These methods are based on the definition of homologous genes within a selected set of genomes, but the increase in evolutionary distance among genomes weakens the evidence of homologous genes. These matters are common unavoidable problems in the comparative genomics research.

We focused on the cyanobacterial genomes to establish an efficient handling method for this problem. The cyanobacteria belong to a group of prokaryotes, which has the ability of oxygenic photosynthesis. The current habitat of cyanobacteria covers most of the planet surface from an extreme environment such as hot springs to water stains on cement walls in city and some species also live on deserts. Some of the reasons for their wide distribution can be related to their long evolutionary history, which dates from 3.5 Ga by geological evidence. The genomes of cyanobacteria are energetically sequenced by Kazusa DNA Institute and Joint Genome Institute, among others. Currently, 16 genome sequences have been published (as of 2005), and this number is the largest for a single group of prokaryotes.

As mentioned above, cyanobacterial evolutionary divergence is so large that the phylogenetic trees generated from molecular evolutionary analyses in 16S rRNA are sometimes inconsistent. In the present study, Gclust database[1], which is designed for analysis of cyanobacteria are used as a source of homologous sequences. We applied approaches of comparative genomics to develop the appropriate data evaluation and to shed light on their complicated evolutionary history of cyanobacteria.

2 Method and Results

Process Overview

In this study, we measure the distance of a pair of clusters instead of genes. Here, 'cluster' is used as defined in the COG database [2], namely, group of homologous genes. Each cluster consists of genes that have high sequence similarities with each other. The relationships among members of a cluster is not always simple, because it contains multiply copied genes ('paralogous genes') in one genome. In other words, gene clusters are a union of homologous and paralogous genes. Therefore, different approaches are needed to process clusters. Consider the homologous gene pair p, q , the distance of genes are measured in a count of genes between gene p and q . We denote this distance c_{pq} . On the other hand, if clusters are constructed from multiple genomes, genes in a cluster P can be represented, for

example, as $P = \{p_{\alpha 1}, p_{\alpha 2}, p_{\beta}, p_{\gamma 1}, p_{\gamma 2}, p_{\gamma 3}, \dots\}$, where Greek subscript indicates genomes ($\alpha, \beta, \gamma, \dots$) and numeral subscript indicates paralogous genes. By fixing genome to α , we get the subset of this cluster $P_{\alpha} = \{p_{\alpha 1}, p_{\alpha 2}\}$. Here, this cluster has two genes in genome α . The distance of a pair of clusters P and Q is defined by the following equation:

$$D_{PQ} = \min\{c_{pq} \mid p \in P, q \in Q\} \quad (1)$$

We made vectors of cluster distances defined by Equation 1 with the threshold of 100. Each vector consists of the distance from a cluster to all other clusters in a genome. These vectors were obtained for all the genomes and stored as distance tables that have information of the alignment of the clusters. These distance tables were used for the analysis.

Gclust Data

The dataset of homologous genes are obtained from the Gclust database. The 16 cyanobacterial genomes were used. Further details about datasets are described at Gclust database site[3]. This database stores the homology of all genes of the genomes in some perspectives and the clustering results by these scores. We took the mapping between gene ID and cluster ID from transient files of the database generation.

Cluster Distance Structures in *Cyanobacteria*

As a result, we selected 4463 conserved clusters from 8980 original clusters in the database. A half of the original clusters are mostly singletons and are not significant. Interestingly, many of the obtained clusters are preferentially located within nearby 100 clusters to both directions. The 16 genomes were classified into at least two groups by the order pattern, and the cluster members of these series are independent from each group.

3 Discussion

The relationships of the homologous genes as usually used are one-to-one, while, in our approach, the relationships of clusters are many-to-many. Despite such relaxation of conditions the observations of the patterns of cluster distance lead to the fact that the gene cluster, which is a group of homologous genes, can be interchangeable with strictly homologous genes. This shows that our method can apply for more applications in comparative genomics. In our approach, the Gclust database is the best suitable because the cluster size was regulated in balance between the sequence homology scores and the numbers of homologous genes in order that a cluster contains at least one gene in all genomes.

Acknowledgment

Computation time was provided by the Super Computer System, Human Genome Center, Institute of Medical Science, University of Tokyo.

References

- [1] Sato, N., Ishikawa, M., Fujiwara, M., and Sonoike, K., Mass identification of chloroplast proteins of endosymbiont origin by phylogenetic profiling based on organism-optimized homologous protein groups, *Genome Informatics*, 16:56–68, 2005.
- [2] Tatusov, R. L., Galperin, M. Y., and Koonin, E. V., The COG database: A tool for genome-scale analysis of proteins functions and evolution, *Nucleic Acids Res.*, 25:3389–3402, 2000.
- [3] <http://gclust.c.u-tokyo.ac.jp/>