Extraction of Organism Groups from Whole Genome Comparisons

Yoshihiro Yamanishi[1] yoshi@kuicr.kyoto-u.ac.jp
Akiyasu C. Yoshizawa[1] acyshzw@kuicr.kyoto-u.ac.jp
Masumi Itoh[1] itoh@kuicr.kyoto-u.ac.jp
Toshiaki Katayama[2] ktym@hgc.jp
Minoru Kanehisa[1] kanehisa@kuicr.kyoto-u.ac.jp

[1] Bioinformatics Center, Institute for Chemical Research, Kyoto University, Gokasho, Uji, Kyoto 611-0011, Japan
[2] Human Genome Center, Institute of Medical Science, University of Tokyo, 4-6-1 Shirokane-dai, Minato-ku, Tokyo 108-8639, Japan

Keywords: orthologous gene, genome tree, phylogenetic profile, independent component analysis

1 Introduction

The availability of a growing number of fully sequenced genomes makes it possible for us to conduct a large-scale comparative genomic research. In recent years, functional prediction and genome tree construction methods based on phylogenetic profiles have been developed. The phylogenetic profile is defined as a bit pattern that encodes the presence or absence of orthologous genes in a set of organisms with fully sequenced genomes, so the phylogenetic profile can be regarded as the evolutionary pattern of gene acquisition and gene loss across organisms. The genome tree methods based on phylogenetic profiles stem from the assumption that evolutionarily correlated organisms have similar gene inheritance patterns. The purpose of this paper is to conduct a whole genome scale phylogenetic analysis in order to extract major organism groups from the comparison of fully sequenced genomes using the independent component analysis, at the same, to identify genes that are characteristic to a certain organism group. In the experiment, we show that 24 major organism groups are extracted from phylogenetic profiles composed of orthologous gene clusters, and our genome tree corresponds to the simplified classification of KEGG organism groups and suggests that two prokaryotic kingdoms, Archaea and Bacteria, are closer to each other than to Eukaryotes on the whole genome level. We also show a possibility for predicting biological functions of hypothetical genes by examining the correspondence between the extent of their affiliation and specific properties of certain organism groups.

2 Material and Method

Ortholog clusters (OCs) are sets of orthologous gene clusters, which are currently developed in the KEGG database [1]. The gene clusters which are identified as cliques in the graph of SSDB (sequence similarity database) are assigned unique identifiers called OC numbers. In this study, we focus on 126 organisms with fully and partially sequenced genomes (including 11 eukaryotes, 99 bacteria and 16 archaea) in the KEGG/GENES and KEGG/SSDB databases as of June 2003. Similarly as phylogenetic profiles, OC profiles are constructed from 17286 OCs (whose gene members exist in at least two organisms) in the 126 organisms. Each OC profile consists of a string of bits, where the presence and absence of an OC are coded as 1 and 0 across organisms.

We applied the independent component analysis (ICA) to the OC profile matrix. The original OC profile matrix of 17286 rows (OCs) and 126 columns (organisms) was converted to a matrix of 17286 rows and 24 columns (independent components). Furthermore, to obtain a hierarchy of organisms (genome tree), we performed a cluster analysis based on the set of 24-variable vectors containing the correlation coefficients between 126 organisms and 24 independent components (ICs).

3 Results and Discussion

Table 1 summarizes the result of assigning each independent component (IC) to a specific organism group, where the numbers IC1 to IC24 were numbered manually in the order of eukaryotes, bacteria, and archaea. The score in each IC enables us to evaluate the extent of affiliation of genes to specific organism groups. Actually we confirmed the validity of the extracted organism group from the viewpoint of biological functions. For instance, high scoring genes in IC3 correspond to DNA-directed RNA polymerase III subunit, high scoring genes in IC20 correspond to photosystem, and high scoring genes in IC21 and IC24 correspond to ribosomal proteins 30S and 50S. These results show a possibility of predicting biological functions of hypothetical genes identified to be characteristic to a certain organism group, by examining their correspondence with specific phenotypes of the organism group.

On the contrary to the genome tree based on the KEGG orthology (KO) [1], the genome tree based on the ortholog cluster (OC) in this study suggests that the Archaea is in proximity to the Bacteria rather than the Eukaryotes, and this result is consistent with the whole genome trees in several literatures. One explanation of the difference is that the OC is a set of orthologs on the whole genome scale, while the KO is a set of the orthologs involved mainly in metabolic pathways.

Table 1: Organism groups extracted by ICA. “E” indicates Eukaryote, “B” indicates Bacteria, “P” indicates Proteobacteria, “F” indicates Firmicutes, and “A” indicates Archaea, respectively.

IC	Organism group	IC	Organism group	IC	Organism group
IC1	E/Animal	IC9	B/P/beta,delta,epsilon	IC17	B/Actinobacteria(1)
IC2	E/Insect, nematode	IC10	B/P/alpha(1)	IC18	B/Actinobacteria(2)
IC3	E/Fungi	IC11	B/P/alpha(2)	IC19	B/Chlamydia
IC4	B/P/gamma(1)	IC12	B/P/alpha(3)	IC20	B/Cyanobacteria
IC5	B/P/gamma(2)	IC13	B/F/Bacillales(1)	IC21	B
IC6	B/P/gamma(3)	IC14	B/F/Bacillales(2)	IC22	A/Euryarchaeota
IC7	B/P/gamma(4)	IC15	B/F/Lactbacteria	IC23	A/Crenarchaeota
IC8	B/P/gamma(5)	IC16	B/F/Clostridia	IC24	A

4 Acknowledgments

This work was supported by grants from the Ministry of Education, Culture, Sports, Science and Technology of Japan, the Japan Society for the Promotion of Science, and the Japan Science and Technology Corporation. The computational resource was provided by the Bioinformatics Center, Institute for Chemical Research, Kyoto University.

References

[1] Kanehisa, M., Goto, S., Kawashima, S., and Nakaya, A., The KEGG databases at GenomeNet, *Nucleic Acids Res.*, 30:42–46, 2002.

[2] Yamanishi, Y., Itoh, M., and Kanehisa, M., Extraction of organism groups from phylogenetic profiles using independent component analysis, *Genome Informatics*, 13:61–70, 2002.