Functional Categorization of Multiple Genomes using KEGG OC in the Genome Indices

Shujiro Okuda[1]
okuda@kuicr.kyoto-u.ac.jp

Akiyasu C. Yoshizawa[1]
acyshzw@kuicr.kyoto-u.ac.jp

Yuki Moriya[1]
moriya@kuicr.kyoto-u.ac.jp

Masumi Itoh[1]
itoh@kuicr.kyoto-u.ac.jp

Toshiaki Katayama[2]
ktym@hgc.jp

Susumu Goto[1]
goto@kuicr.kyoto-u.ac.jp

Minoru Kanehisa[12]
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: KEGG, Ortholog Cluster (OC), gene ontology (GO)

1 Introduction

The increasing availability of complete and draft genomic sequences enables us to perform large scale comparative studies of the genomes. Now about 200 genomes are available and we can computationally analyze them from a variety of viewpoints. In order to facilitate such analyses and to identify characteristic properties in each genome, we have been constructing the database named Genome Indices[3]. It stores various biological indices in multiple genomes, which can be easily accessed through the web-interface. We previously reported the collection of basic statistical values computed from genomic sequences and also from biological pathways. We are now implementing an additional feature of the Genome Indices, which is the statistics on functional categories of genes. The functional categories are taken from the EC numbers, KEGG Orthology (KO), Clusters of Orthologous Groups (COG) and Gene Ontology (GO)[1]. The assignment of genes in these resources are matched to the KEGG Ortholog Cluster (OC), which is an automatic grouping of orthologous genes in the complete genomes. Here we report how genomes can be characterized by gene contents of different categories, in particular the GO categories measured by the OC grouping.

2 Materials and Method

2.1 Ortholog Clusters

The Ortholog Clusters (OC) were obtained from the KEGG database at GenomeNet[2]. The OCs we used are the sets of orthologous gene clusters based on the KEGG SSDB database storing the sequence similarities by the Smith-Waterman algorithm. The current version of OCs consists of 611,655 genes from 179 genomes including 17 eukaryotes, 18 archaea and 144 bacteria. These genes are decomposed into 124,404 clusters including singletons, among which 33,390 clusters represent ortholg relations in different organism groups.

2.2 Gene Ontology

We used Gene Ontology (GO)[1] for functional categorization of genes. GO contains biological terms organized in the directed acyclic graph. We used GO Slim, which is the set of slimmed down terms that are mapped to original GO terms. The GO Slim was created with the perl script map2slim.pl provided by the GO website. More than 18,000 terms were slimmed down to 138 terms. We also obtained GO annotations from the GO website. Gene IDs in the annotation files were mapped to genes from KEGG GENES[2].

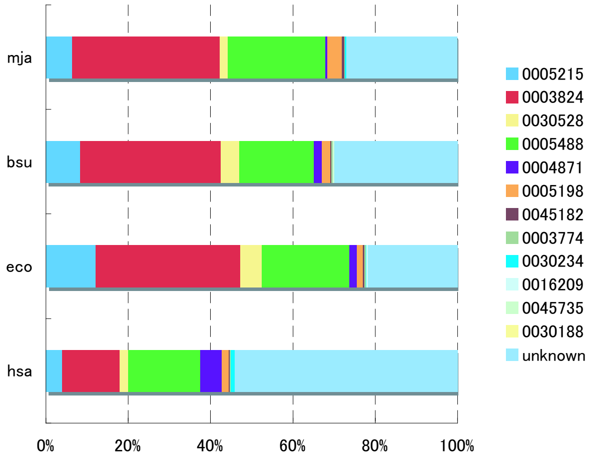

Figure 1: Distribution of functional categories among four representative species in different taxa.

2.3 Assignment

We extracted genes assigned to GO from each OC and a higher level GO Slim term was regarded as the representative functional category of the OC when 90 % of the genes share the term. Multiple terms can be assigned to a certain OC, because a gene can be mapped to more than one GO term. As a result 28,702 OCs could be annotated by the GO Slim functional categories.

3 Results and Discussion

Figure 1 shows the distribution of the GO Slim functional categories in each genome. Bacteria and archaea showed high coverage of functional assignments, while in eukaryotes including human roughly a half of the genes are indicated as function unknown. The most abundant functional categories are catalytic activity (ID: 0003824) and binding (ID: 0005488). In the near future, we will add the other functional categories using the resources such as KO. These indices will characterize genomes from different perspectives and help to understand evolutionary and functional relationships of different genomes.

4 Acknowledgments

This work was supported by grants from the Ministry of Education, Culture, Sports, Science and Technology, 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] Harris, M. A., Clark, J., Ireland, A., Lomax, J., Ashburner, M., Foulger, R., Eilbeck, K., Lewis, S., Marshall, B., Mungall, C., *et al.* The Gene Ontology (GO) database and informatics resource, *Nucleic. Acid. Res.*, 32:D258-61, 2004.

[2] Kanehisa, M., Goto, S., Kawashima, S,. Okuno, Y,. and Hattori, M. The KEGG resource for deciphering the genome, *Nucleic. Acid. Res.*, 32:D277-280, 2004.

[3] http://gi.kuicr.kyoto-u.ac.jp/