Automatic generation of KEGG OC (Ortholog Cluster) and its assignment to draft genomes

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

Toshiaki Katayama[2]
ktym@hgc.jp

Akihiro Nakaya[3]
nakaya@k.u-tokyo.ac.jp

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

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

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

Minoru Kanehisa[1]
kanehisa@kuicr.kyoto-u.ac.jp

1 Bioinformatics Center, Institute for Chemical Research, Kyoto, Gokasho, Uji, Kyoto 611-0011, Japan
2 Human Genome Center, Institute of Medical Science, University of Tokyo4-6-1 Shirokanedai, Minato-ku, Tokyo 108-8639, Japan
3 Department of Computational Biology, University of Tokyo, 5-1-5, Kashiwanoha, Kashiwa-shi, Chiba, 277-8561

Keywords: ortholog cluster, KEGG, genome annotation

1 Introduction

As the number of sequenced genomes are rapidly growing, a method for automatic generation of orthologous gene clusters is needed. However, it is computationally hard to cluster a large number of genes at once. To address this problem, we have developed a heuristic method to assign gene groups from closely related organisms to an ortholog cluster in a bottom-up approach. In this method, we consider each gene subgroup as a representative gene and find their correspondence using bi-directional best hit (BBH) relations obtained from the KEGG SSDB database which stores all-vs-all Smith-Waterman similarity scores [1]. We have clustered all the genes in the KEGG GENES database [1] to generate KEGG Ortholog Clusters (OCs) which represent various aspects of the protein universe. As an application of KEGG OC, we have performed automatic gene assignment of the draft genomes which are not yet included in KEGG. Our method provides an efficient way for rapid annotation of the genes of newly sequenced organisms.

2 Methods and Results

2.1 Generation of KEGG Ortholog Clusters

We have applied a graph-based algorithm for finding quasi-cliques to detect gene clusters. Firstly, we have clustered paralogous genes based on the Smith-Waterman scores between all pair of genes in each organism. Then, each pair of paralog clusters from taxonomically close organisms are merged into an orthologous gene cluster according to the number of BBH (bi-directional best hit) relations. The merged cluster is divided into quasi-cliques by a heuristic method [2] to suppress the cluster size and to prevent remotely related genes from being accumulated. These steps are iterated along the phylogenetic tree until corresponding clusters from all organisms are merged. We have used 611,655 genes from 178 organisms from the KEGG GENES database and obtained 124,404 OCs as a result. The distribution of the number of genes in each OC approximately follows the power law, where 78,263 clusters only contain one gene (singletons) and 928 clusters contain over 100 genes. Among these, 38,019 clusters included at least two organisms.

2.2 Evaluation of KEGG OC assignments

Next, we have attempted to assign genes to the OCs by simply using their best hit from the BLAST search. To test the accuracy of this method, we first randomly selected 5,000 genes as query genes. For each gene in turn, we removed the organism containing the selected gene from KEGG OC. Then we assigned each gene to the OC containing the best-hit gene. We also have tested genes based on organism, including *Homo*

sapiens (hsa) and *Bacillus subtilis* (bsu). Table 1 shows the results including the percentages of genes that were successfully and unsuccessfully assigned, followed by the ratio of correct and incorrect assignment for each test. About 80% of the randomly selected genes were assigned to OCs of which about 90% were re-assigned to their original OCs correctly. In this table, the “Could not assign” category refers to those genes that could not be assigned, and "Correct" in this category indicates the number of genes which were contained in the clusters unique to a particular organism. As the number of *Bacillus* genomes are already sequenced and included in KEGG, it is easier to find similar genes in the datasets for 'bsu' as opposed to 'hsa', which leads to the better accuracy of the assignments in 'bsu'. On the contrary, the accuracy for the human datasets can be improved when there are more eukaryotic genomes available in KEGG.

Table 1: Accuracy of the gene annotation by OC assignment method

Category	Random 5000 seq.	*Homo sapiens* (16468)	*Bacillus subtilis* (4106)
Assigned proteins	4807 (80.5%)	15114 (91.8%)	3540 (86.2%)
Correct	3678 (73.6%)	11686 (71.0%)	3277 (79.8%)
Incorrect	337 (6.7%)	3428 (20.8%)	263 (6.4%)
Could not assign	974 (19.5%)	1354 (8.2%)	566 (13.8%)
Correct	789 (15.8%)	1250 (7.6%)	497 (12.1%)
Incorrect	185 (3.7%)	104 (0.6%)	69 (1.7%)
Over all accuracy	4467 (89.3%)	12936 (78.6%)	3774 (91.9%)

3 Discussion

Finding orthologs is an effective way to estimate the biological functions for unannotated genes. For this purpose, we have developed an automated way for generating ortholog clusters for all organisms, which is used regularly to update our data. However, calculating ortholog clusters requires enormous computational power even when employing our method. Thus, it is not practical to generate ortholog clusters for the draft genomes because gene prediction is frequently revised. The proposed method enables us to assign genes to OCs without clustering, so that now we can assign OCs for a whole set of predicted proteins of 26 eukaryotic draft genomes obtained from Ensembl, TIGR, Broad Inst., Sanger Inst. and JGI. We will summarize the result of our comprehensive assignment for the draft genomes in our poster.

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] Kanehisa, M. Goto, S. Kawashima, S. Okuno, Y. Hattori, M., The KEGG resource for deciphering the genome, *Nucleic Acids Res.*, 32 Database issue:D277-D280, 2004.

[2] Matsuda, H. Ishihara, T. and Hashimoto, A., Classifting Molecular Sequences using a Linkage Graph with Their Pairwise Similarities, *Theor. Comput. Sci.*, 210:305-325, 1999.