CyanoClust: Protein Cluster Database for Comparative genomics of Cyanobacteria and Plastids

Naoki Sato¹
naokisat@bio.c.u-tokyo.ac.jp

Naobumi V. Sasaki¹
naobumi@bio.c.u-tokyo.ac.jp

¹ Department of Life Sciences, Graduate School of Arts and Sciences, University of Tokyo, 3-8-1 Komaba, Meguro-ku, Tokyo, 153-8902, Japan

Keywords: homology cluster database, comparative genomics, cyanobacteria, plastid

1 Introduction

Chloroplasts are the sites of photosynthesis within the cell of land plants and algae. In non-photosynthetic tissues of plants, they are called plastids. Various lines of evidence, such as similarity of photosynthetic machineries and photosynthetic metabolic pathways as well as fossil records, suggested that the cyanobacteria are related to the origin of chloroplasts [1]. Now, cyanobacteria exhibit ecologically and morphologically adapted to wide ecological spectrum from Antarctica to the desert and even in some hot springs. Although their evolutionary history is complicated, the clade structure and the lineage of cyanobacteria have been revealed by careful molecular phylogenetic analyses. These findings also encourage the efforts to deciphering the endosymbiosis process of the chloroplasts. Nevertheless, instead of the intuitive idea that the plastids and the host plant cells were co-evolved, actual evolutionary relationship between the host cells and the plastids still remains complex as seen in the secondary symbiosis [2]. To study the evolution of plastids, there are some useful platforms of comparative genomic analyses such as ChloroplastDB [4]. But resources for comparative genomics of plastids and cyanobacteria are limited. Here we developed CyanoClust database web server for providing reliable orthologs over many cyanobacteria and plastids, or just various cyanobacteria, which is also used as a platform of comparative genome analysis. For the comparison of orthologs among loosely related organisms, the Gclust software, which was developed in our laboratory [3] was used for the construction of the database.

2 Methods and Results

2.1 Database content and Organization

CyanoClust (version 2009-01) contains protein homology information in 34 cyanobacteria, 54 plastids, three photosynthetic bacteria, and one chromatophore of Paulinella. In addition to these photosynthetic organisms, two non-photosynthetic bacteria Escherichia coil K-12 and Bacillus subtilis Marburg 168 were also included, which were expected to serve as outgroups. The summary of the database building process is shown in Fig. 1. We used the Gclust software, which is based on the Entropy Optimized Organism Count (EOOC) method [3], for automatic clustering to obtain a homolog cluster table.

In total, 137988 proteins were classified into 30793 clusters, including 19412 singleton clusters for which no orthologs were detected in any of the other genomes. Each cluster is accessed by a serial ID, and each entry of cluster contains the list of gene IDs. In addition, assignment to one of the large conserved synteny groups that we call virtual linkage groups (VLG) is indicated with a probable consensus annotation, if available. The ‘gene’ tables and ‘homology cluster’ table are mainly used in the CyanoClust web system. The fields emphasized by slant bold in these tables (Fig. 1) were used as search keys in the web interface.

Figure 1: Concept schema of the database.
PK, primary key; FK, foreign key.
2.2 Cyanoclust interface
The web system of the Cyanoclust are driven by CGI scripts. Most search tasks can be performed in the ‘Basic search’ page and the ‘BLAST search’ page. The ‘Basic search’ page accepts a search query by individual gene name or annotations, gene existence in specified organism groups or species, or directly specified cluster ID.

By submitting a query, a ‘Cluster ID’ page is displayed, which contains information on cluster ID, sequence ID and the representative annotation of each cluster. In keyword search, the keyword hit is marked by red color. Then, one can access the clusters or protein sequences of one’s interest. Users can fetch all sequences in multi-FASTA format for further analysis. In addition, searching from an arbitrary protein sequence is available from the ‘BLAST search’ page, which is available by following the ‘BLAST search’ link on top of the basic search page. Cyanoclust web system also implements convenient phylogenetic analysis tool using JalView software (Fig. 2). With these interfaces, Cyanoclust web system is used as a platform of comparative genomic analysis.

3 Discussion
Current Cyanoclust database does not modify the annotations given for each protein in the original database and this field just inherits the gene annotation of the top entry of the gene list. An idea of future enhancement is to give a well curated annotation to each cluster. But this will only be possible after extensive efforts of manual curation. Although extensive, the complete curated annotation will be possible by combining all available experimental data through the orthologous relationship defined by the Cyanoclust, we will be able to give a more reliable annotation to each conserved cluster.

Acknowledgements
This study was supported in part by Grants-in-Aid for Scientific Research from the MEXT, Japan (nos. 17018010, 18017005, 20017006) and the Global COE program “From Earth to Earths” to NS, and a Grant-in-Aid from the JSPS (no. 2011425) to NVS.

References