LABORATORY OF GENOME INFORMATICS

Assistant Professor
UCHIYAMA, Ikuo

Postdoctoral Fellow: *CHIBA, Hirokazu*

The accumulation of biological data has recently been accelerated by various high-throughput "omics" technologies including genomics, transcriptomics, and proteomics. The field of genome informatics is aimed at utilizing this data, or finding the principles behind the data, for understanding complex living systems by integrating the data with current biological knowledge using various computational techniques. In this laboratory we focus on developing computational methods and tools for comparative genome analysis, which is a useful approach for finding functional or evolutionary clues to interpreting the genomic information of various species. The current focus of our research is on the comparative analysis of microbial genomes, the number of which has been increasing rapidly. Since different types of information about biological functions and evolutionary processes can be extracted from comparisons of genomes at different evolutionary distances, we are developing methods to conduct comparative analyses not only of distantly related but also of closely related genomes.

I. Microbial genome database for comparative analysis

The microbial genome database for comparative analysis (MBGD; URL http://mbgd.genome.ad.jp/) is a comprehensive platform for microbial comparative genomics. The central function of MBGD is to create orthologous groups among multiple genomes from precomputed all-against-all similarity relationships using the DomClust algorithm (see Section II below). By means of this algorithm, MBGD not only provides comprehensive orthologous groups among the latest genomic data available, but also allows users to create their own ortholog groups on the fly using a specified set of organisms. In addition, MBGD also provides MyMBGD mode, which allows users to add their own genomes to MBGD.

The database now contains more than 2000 published genomes including 34 eukaryotic microbes and 4 multicellular organisms. This year, we substantially enhanced the database functionality to explore the diversity of microbial genomes. The enhancement includes: 1) To efficiently explore the diversity of the microbial genomic data, MBGD now provides summary pages for pre-calculated ortholog tables among various taxonomic groups; 2) For some closely related taxa, MBGD also provides the conserved synteny information (core genome alignment) pre-calculated using the CoreAligner program (see Section III below); 3) An efficient incremental updating procedure was implemented to create extended ortholog tables by adding additional genomes to the default ortholog table generated from the representative set of genomes (see Section II below).

II. Improvement of the methods for constructing orthologous groups among multiple genomes

As a core technology of our comparative genomics tools, we have developed a rapid automated method of ortholog grouping, named DomClust, which is effective enough to allow the comparison of many genomes simultaneously. The method classifies genes based on a hierarchical clustering algorithm using all-against-all similarity data as an input, but it also detects domain fusion or fission events and splits clusters into domains if required. As a result, the procedure can split genes into the domains minimally required for ortholog grouping.

Although DomClust can rapidly construct orthologous groups, its classification quality has room for improvement since it uses UPGMA like clustering based on pairwise sequence alignment. We are now developing a procedure to refine the DomClust classification based on multiple sequence alignments and phylogenetic trees. Since DomClust constructs orthologous groups at the domain level, the procedure refines not only sequence grouping but also domain splitting. We developed a simple scoring system to identify the most plausible domain boundary or to merge multiple domains into one group by maximizing the sum of the alignment scores of domains under this scoring system (Figure 1). We tested the scoring system using the COG database as a reference and confirmed that changes that increase the score generally improve the agreement with the COG clusters. After determining domain boundaries, our procedure also checks whether the resulting group should be split into subgroups by considering species overlaps between subclusters in the phylogenetic tree (Figure 1).

We are also developing a method to update the clustering result incrementally, by which we can add new genomes to a reference set of ortholog groups that is constructed from a representative set of published genomes available in the MBGD server. This approach allows us to conduct further large-scale comparative genomics based on orthologous groups among thousands of genomic sequences. We are also trying to extend the algorithm for handling metagenomic data. To infer the taxonomic position of the source organism

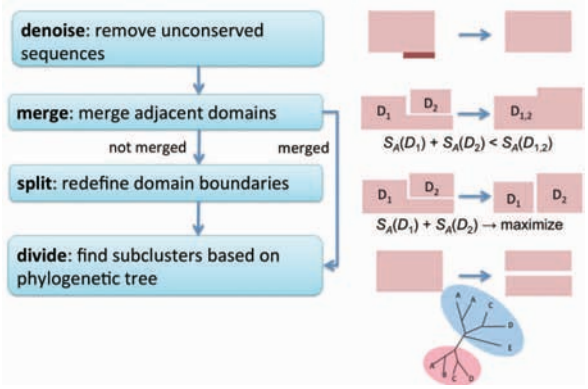

Figure 1. Pipeline for ortholog group refinement. The pipeline consists of several modules including *denoise*, *merge*, *split* and *divide*. Merge and split aim at improving the domain boundary based on the sequence alignment whereas divide aims at improving the subgrouping based on the phylogenetic tree.

of each metagenomic sequence, we have developed a method to map each tree node of the hierarchical clustering tree generated by the DomClust algorithm onto a taxonomic tree node.

III. Identification of the core structure conserved among moderately related microbial genomes

Horizontal gene transfers (HGT) have played a significant role in prokaryotic genome evolution, and the genes constituting a prokaryotic genome appear to be divided into two classes: a "core gene pool" that comprises intrinsic genes encoding the proteins of basic cellular functions, and a "flexible gene pool" that comprises HGT-acquired genes encoding proteins which function under particular conditions. Thus identification of the set of intrinsically conserved genes, or the genomic core, among a taxonomic group is crucial for understanding prokaryotic diversity and evolution.

Typically "core genome" is defined as a set of genes that are conserved among all the genomes belonging to the given species (called the "universal core"). However, this definition can be too strict since it allows no exceptions. We consider the core structure of related genomes as a set of sufficiently long segments in which gene orders are conserved among multiple genomes so that they are likely to have been inherited mainly through vertical transfer (called "syntenic core"). To find such core structures, we developed a method named CoreAligner, which finds the order of orthologous groups that maximally retains the conserved gene orders.

We systematically applied our method to bacterial taxa (family, genus and species) that contain a sufficient number of completed genomes stored in MBGD. As a result, we can generally obtain more core genes as syntenic core rather than universal core genes, except for some taxonomic groups that have poor syntenic conservation. Moreover, typically the number of syntenic core is more stable than universal core when the number of genomes in the given taxa increases.

The core genome data calculated for various prokaryotic taxa is now available as part of the MBGD database (Figure 2).

Figure 2. Core genome alignment viewer in MBGD showing the core alignment of the family *Bacillaceae*.

IV. Development of a workbench for comparative genomics

We are developing a comparative genomics workbench named RECOG (Research Environment for COmparative Genomics), which aims to extend the current MBGD system by incorporating more advanced analysis functionalities including phylogenetic pattern analysis, the ingroup/outgroup distinction in ortholog grouping and the core structure extraction among related genomes. The entire RECOG system employs client-server architecture: the server program is based on the MBGD server and contains the database construction protocol used in MBGD so that users can install the server on their local machines to analyze their own genomic data, whereas the client program is a Java application that runs on a local machine by receiving data from any available RECOG server including the public MBGD server.

The central function of RECOG is to display and manipulate a large-scale ortholog table. The ortholog table viewer is a spreadsheet like viewer that can display the entire ortholog table, containing more than a thousand genomes. Using the zoom in/out function, it can display the entire table or a section of the main table with more detailed information. In RECOG, several basic operations on the ortholog table are defined, which are classified into categories such as filtering, sorting, and coloring, and various comparative analyses can be done by combining these basic operations, such as "Neighborhood gene clustering" and "Phylogenetic pattern clustering." In addition, RECOG allows the user to input arbitrary gene properties such as sequence length, nucleotide/amino acid contents and functional classes, and compared these properties among orthologs in various genomes.

Publication List

〔Original papers〕

- Takami, H., Noguchi, H., Takaki, Y., Uchiyama, I., ,Toyoda, A., Nishi, S., Chee, G-J., Arai, W., Nunoura, T., Itoh, T., Hattori, M., and Takai, K. (2012). A deeply branching thermophilic bacterium with an ancient acetyl-CoA pathway dominates a subsurface ecosystem. PLoS ONE *7*, e30559.
- Yahara, K., Kawai, M., Furuta, Y., Takahashi, N., Handa, N., Tsuru, T., Oshima, K., Yoshida, M., Azuma, T., Hattori, M., Uchiyama, I., and Kobayashi, I. (2012). Genome-wide survey of mutual homologous recombination in highly sexual bacterial species. Genome Biol. Evol. *4*, 628-640.

〔Original paper (E-publication ahead of print)〕

- Uchiyama, I., Mihara, M., Nishide, H., and Chiba, H. MBGD update 2013: the microbial genome database for exploring the diversity of microbial world. Nucleic Acids Res. 2012 Oct 30.