LABORATORY OF GENOME INFORMATICS

Assistant Professor
UCHIYAMA, Ikuo

Postdoctoral Fellow: *KAWAI, Mikihiko*

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 these 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) 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. The latter feature is especially useful when the user's interest is focused on some taxonomically related organisms.

We have continued to develop and enhance the database functionality. Each orthologous group entry is assigned functional annotation and external database links that are a summarization of the information assigned to the individual genes belonging to that group. Phenotypic properties of each genome are stored and can be used for specifying a set of genomes for phylogenetic pattern analysis. MyMBGD mode, which allows users to add their own genomes to MBGD, accepts raw genomic sequences without any annotation. The database now contains well over 1000 published genomes including 23 eukaryotic microbes and 4 multicellular organisms.

MBGD is available at http://mbgd.genome.ad.jp/.

II. Enhancement of the algorithm for identifying 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.

We are continuing to improve the algorithm. To combine closely related genome comparison with distantly related genome comparison, the algorithm accepts ingroup/outgroup specification for each input genome, and considers taxonomic information partially during ortholog grouping. The resulting table has a nested structure when a duplication event occurs within the ingroup lineage. To improve the scalability of the algorithm for comparison of thousands of genomic sequences, we have developed an efficient method to update the clustering result incrementally. This method is especially useful for a user who wants to compare his/her own original genomes with the published genomes available in the MBGD server, where the orthologous relationship among the published genomes has already been calculated and is available.

We are also extending the algorithm for handling metagenomic data. In contrast to usual comparative genome analyses, in metagenomic analysis, the source organism of each metagenomic sequence is not known; instead, taxonomic position of the source organism of each metagenomic sequence should be inferred. For this purpose, we have extended the DomClust algorithm to infer taxonomic position for each metagenomic sequence by mapping 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. The identification of the set of intrinsically conserved genes, or the genomic core, among a taxonomic group is crucial not only for establishing the identity of each taxonomic group, but also for understanding prokaryotic diversity and evolution. 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. We developed a method for aligning conserved regions of multiple genomes, which finds the order of pre-identified orthologous groups that retains to the greatest possible extent the conserved gene orders.

We are now expanding our analysis to more diverged

bacterial families to examine generality of our approach. We are also developing an enhanced algorithm that can incorporate phylogenetic relationships among input genomes.

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 (Figure 1). 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. For example, “Neighborhood gene clustering” identifies a set of genes that are located in the vicinity of each other in both the ortholog table and the genomic sequence, and assigns the same color to each group. “Phylogenetic pattern clustering” performs hierarchical cluster analysis based on the dissimilarity between phylogenetic patterns, and reorders the ortholog table according to the clustering result. 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.

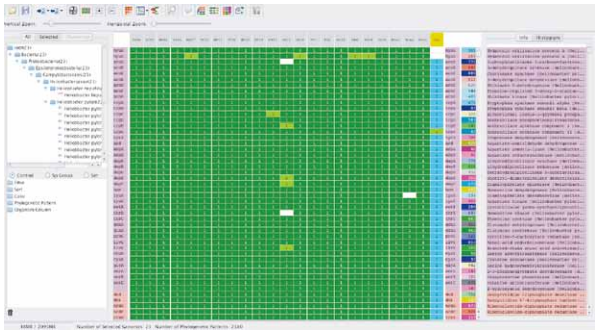

Figure 1. The RECOG client program showing orthologous relationships among 22 *Helicobacter pylori* strains including four Japanese strains determined in our project. *Helicobacter hepaticus* is also added as an outgroup species in the rightmost column.

V. Comparative genomics of *Helicobacter pylori*

Helicobacter pylori is a major pathogen in human gastric cancer and it is known that the East Asian strains of *H. pylori* have a stronger subtype of a major virulence factor, CagA protein, than Western strains. In collaboration with Dr. Kobayashi (Univ. Tokyo) and other researchers, we have determined the complete genomic sequences of four *H. pylori* strains isolated from Japanese patients and compared them with other published *H. pylori* genomes. Using the RECOG system and other tools (Figure 1), we tried to identify characteristic genomic features of the East Asian strains from various points of view and infer evolutionary processes and mechanisms that are related to the evolution of *H. pylori*. As a result, we were able to identify several genes that characterize the East Asian strains. For example, almost all of the molybdenum-related genes, which are related to the catalysis of two-electron redox reactions, are disrupted specifically in the East Asian strains (Figure 2).

In addition, we found that some outer membrane proteins that are specifically duplicated in the East Asian strains are located at the boundary of the chromosomal inversion identified between the East Asian and other strains. After detailed examination of the boundary of these and other chromosomal inversions identified among *H. pylori* genomes, we were able to find a novel mechanism of genome evolution named DNA duplication associated with inversion.

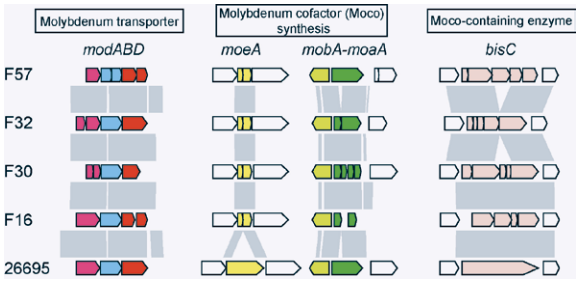

Figure 2. Systematic disruption of the molybdenum-related genes among the Japanese strains (F57, F32, F30, F16). The strain 26695 is a European strain, which has intact genes.

Publication List

〔Original paper〕

● Uchiyama, I., Higuchi, T., and Kawai, M. (2010). MBGD update 2010: toward a comprehensive resource for exploring microbial genome diversity. Nucleic Acids Res. *38*, D361-D365.