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.

This year, we have developed the following advanced functionalities: 1) enhanced assignment of functional annotation, including external database links to each orthologous group, 2) an interface for choosing a set of genomes based on phenotypic properties (Figure1), 3) the addition of more eukaryotic microbial genomes (fungi and protists) and some higher eukaryotes as references, 4) enhancement of the MyMBGD mode, which allows users to add their own genomes to MBGD and now accepts raw genomic sequences without any annotation (in such cases it runs a gene-finding procedure before identifying the orthologs). Some analysis functions, such as the function to find orthologs with similar phylogenetic patterns, have also been improved. The database now contains around 1000 published genomes including 16 eukaryotic microbes and 4 multicellular organisms.

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

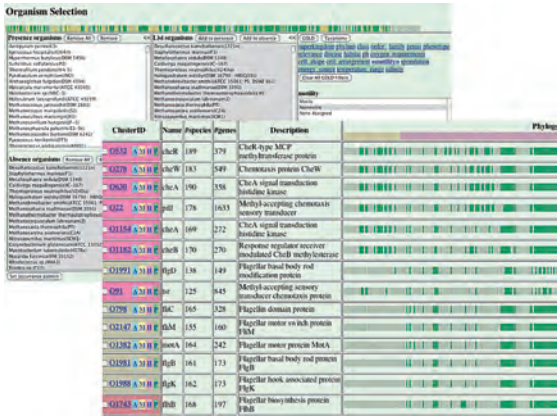

Figure 1. The interface for organism selection in MBGD where a phylogenetic pattern related to "motility" is specified (top-left) and the result of the pattern search (bottom-right).

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. Partial taxonomic information can be incorporated into the DomClust algorithm by specifying ingroup/outgroup for each input genome. The resulting table has a nested structure when a duplication event occurs within the ingroup lineage. We are also trying to modify the algorithm for utilizing metagenomic analysis. In addition, to further improve scalability for comparison of thousands of genomic sequences, we are now developing an efficient method to update the clustering result incrementally.

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.

The program, named CoreAligner, was successfully applied to the genome sequences of two major bacterial families, *Bacillaceae* and *Enterobacteriaceae*, for identifying the core structures comprising 1438 and 2125 orthologous groups, respectively. 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 main window of the RECOG client consists of a taxonomic tree viewer (left), an ortholog table/phylogenetic pattern map viewer (center) and a gene information viewer (right). Users can choose a set of genomes using the taxonomic tree viewer and run the DomClust program to identify orthologous groups. The result is displayed in the ortholog table viewer, where users can see the entire picture of the phylogenetic pattern map as well as the full details of the ortholog table using semantic zooming functionality.

This year, we added several new functions to the RECOG system to enhance its usability for comparative analysis. One of the new features is gene property analysis, in which users can define property values for each gene and compare these values among orthologous genes on the ortholog table viewer (Figure 2).

Figure 2. The phylogenetic pattern viewer comparing ten *H. pylori* strains (see section V), where each cell is colored according to G+C content at the third codon positions using the gene property analysis functionality.

V. Comparative genomics of *Helicobacter pylori*

Helicobacter pylori is a major pathogen in human gastric cancer and it is known that East Asian strains of *H. pylori* have a more toxic CagA protein, a major virulence factor, 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 published *H. pylori* genomes. Although all four genomes are clearly classified into the East Asian group, they exhibit substantial polymorphisms in genomic structure (Figure 3). We are trying to identify genomic features specific to East Asian strains and infer evolutionary processes and mechanisms that are related to the evolution of *H. pylori*.

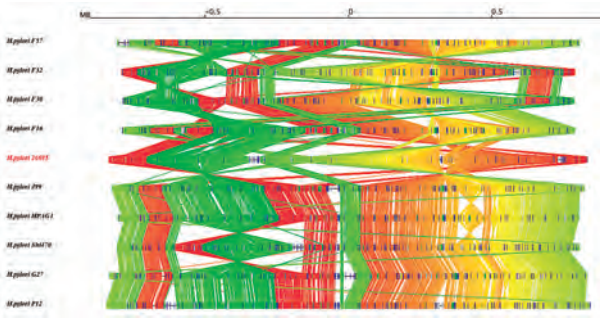

Figure 3. Comparative genome map among *H. pylori* strains. Genomic core structure is identified by CoreAligner and orthologous genes are connected with lines. Colors are assigned according to the gene order in one strain (26695).

Publication List

〔Original papers〕

- Baba, T., Kuwahara-Arai, K., Uchiyama, I., Takeuchi, F., Ito, T., and Hiramatsu, K. (2009). Complete genome sequence determination of a *Macrococcus caseolyticus* strain JSCS5402 reflecting the ancestral genome of the human pathogenic staphylococci. J. Bacteriol. *191*, 1180-1190.
- Watanabe, S., Ito, T., Sasaki, T., Li, S., Uchiyama, I., Kishii, K., Kikuchi, K., Skov, R.L., and Hiramatsu, K. (2009). Genetic diversity of staphylocoagulase genes (coa): insight into the evolution of variable chromosomal virulence factors in *Staphylococcus aureus*. PLoS One *4*, e5714.

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

- Uchiyama, I., Higuchi, T., and Kawai, M. MBGD update 2010: toward a comprehensive resource for exploring microbial genome diversity. Nucleic Acids Res. 2009, Nov. 11.