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 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, but also allows users to create their own ortholog groups on the fly using a specified set of organisms. MBGD also has pre-calculated ortholog tables for each major taxonomic group, and provides several views to display the entire picture of each ortholog table. For some closely related taxa, MBGD provides the conserved synteny information calculated using the CoreAligner program (see Section III below). In addition, MBGD provides MyMBGD mode, which allows users to add their own genomes to MBGD.

Because of the rapid increase in microbial genome data owing to next generation sequencing technology, it becomes more challenging to maintain high quality orthology relationships while allowing users to utilize the latest genomic data available. Since recently accumulating sequences are mostly draft genome data, MBGD now stores them and allows users to incorporate them into a user specific ortholog database through the MyMBGD functionality. In MyMBGD, draft genome data as well as user genome data are incorporated into an existing ortholog table created from complete genome data in an incremental manner. In addition, to provide high quality orthology relationships, the standard ortholog table, which is first created by DomClust, is now refined using the DomRefine program (see Section II below).

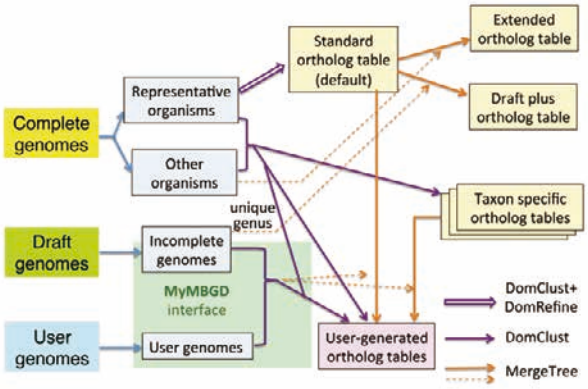

Figure 1. Overview of the data construction procedure in MBGD

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

As a core technology of our comparative genomics tools, we 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 at the domain level, its classification quality has room for improvement since it is based on pairwise sequence alignment. We developed a procedure to refine the DomClust classification based on multiple sequence alignments. The method is based on a scoring scheme, the domain-specific sum-of-pairs (DSP) score, which evaluates ortholog clustering results at the domain level as the sum total of domain-level alignment scores. We developed a refinement pipeline to improve domain-level clustering, DomRefine, by optimizing the DSP score. We applied DomRefine to domain-based ortholog groups created by DomClust using a dataset obtained from the MBGD database, and evaluated the results using COG and TIGRFAMs as the reference data. Thus, we observed that the agreement between the resulting classification and the classifications in the reference databases is improved in the refinement pipeline. Moreover, the refined classification showed better agreement than the classifications in the eggNOG databases when TIGRFAMs was used as the reference database.

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: core and accessory. The core gene set comprises intrinsic genes encoding the proteins of basic cellular functions, whereas the accessory gene set comprises HGT-acquired genes encoding proteins which function under particular conditions. 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, and developed a method named CoreAligner to find such structures. We systematically applied the method to bacterial taxa to define their core gene sets, and are now trying to utilize this information to characterize novel genomic and metagenomic data.

IV. Development of a workbench for comparative genomics

We have developed 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. 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. 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. In addition, RECOG allows the user to input arbitrary gene properties and compare these properties among orthologs in various genomes. We continue to develop the system and apply it to various genome comparison studies under collaborative research projects.

V. Ortholog data representation using Semantic Web technology to integrate various microbial databases

Orthology is a key to integrate knowledge about various organisms through comparative analysis. Moreover, presence/absence of orthologs in each genome can be an important clue to understand the relationship between gene functions and species phenotype/habitat.

We have constructed an ortholog database using Semantic Web technology, aiming at the integration of numerous genomic data and various types of biological information. To formalize the structure of the ortholog information in the Semantic Web, we have constructed the Ortholog Ontology (OrthO). While the OrthO is a compact ontology for general use, it is designed to be extended to the description of database-specific concepts. On the basis of OrthO, we described the ortholog information from MBGD in the form of Resource Description Framework (RDF) and made it available through the SPARQL endpoint, which accepts arbitrary queries specified by users. In this framework based on the OrthO, the biological data of different organisms can be integrated using the ortholog information as a hub.

This is part of the MicrobeDB project, a collaborative project under the National Bioscience Database Center.

VI. Identification of mobile genes and its application to characterizing *H. pylori* pan-genome repertoire

Gene contents of the same bacterial species can have great variation and thus the whole repertoire of genes in a bacterial species, termed the pan-genome, can be very large. We analyzed the pan-genome identified among 30 strains of the human gastric pathogen *Helicobacter pylori* isolated from various phylogeographical groups. For this purpose, we developed a method to define mobility of genes against the reference coordinate determined by the core alignment created by CoreAligner, and classified each accessory gene into mobility classes. We also identified co-occurring gene clusters using phylogenetic pattern clustering combined with neighboring gene clustering implemented in the RECOG system. On the basis of these analyses, we identified several gene clusters conserved among *H. pylori* strains that were characterized as mobile or non-mobile. This work is in collaboration with Prof. Kobayashi, Univ. Tokyo.

Publication List

〔Original papers〕

- Chiba, H., and Uchiyama, I. (2014). Improvement of domain-level ortholog clustering by optimizing domain-specific sum-of-pairs score, BMC Bioinformatics *15*, 148.
- Kawai, M., Futagami, T., Toyoda, A., Takaki, Y., Nishi, S., Hori, S., Arai, W., Tsubouchi, T., Morono, Y., Uchiyama, I., Ito, T., Fujiyama. A., Inagaki, F. and Takami, H. (2014). High frequency of phylogenetically diverse reductive dehalogenase-homologous genes in deep subseafloor sedimentary metagenomes. Front. Microbiol. *5*, 80.
- Matsui, H., Takahashi, T., Murayama, S.Y., Uchiyama, I., Yamaguchi, K., Shigenobu, S., Matsumoto, T., Kawakubo, M., Horiuchi, K., Ota, H., Osaki, T., Kamiya, S., Smet, A., Flahou, B., Ducatelle, R., Haesebrouck, F., Takahashi, S., Nakamura, S., and Nakamura, M. (2014). Development of New PCR Primers by Comparative Genomics for the Detection of *Helicobacter suis* in Gastric Biopsy Specimens. Helicobacter *19*, 260-271.
- Toyota, K., Kato, Y., Miyakawa, H., Yatsu, R., Mizutani, T., Ogino, Y., Miyagawa, S., Watanabe, H., Nishide, H., Uchiyama, I., Tatarazako, N., and Iguchi, T. (2014). Molecular impact of juvenile hormone agonists on neonatal *Daphnia magna*. Appl. Toxicol. *34*, 537-544.

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

- Uchiyama, I., Mihara, M., Nishide, H., and Chiba, H. MBGD update 2015: microbial genome database for flexible ortholog analysis utilizing a diverse set of genomic data. Nucleic Acids Res. 2014 Nov 14.