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 program combined with the DomRefine program (see Section II below). By means of these programs, 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. Moreover, MBGD now stores recently accumulating draft genome data, and allows users to incorporate them into a user specific ortholog database through the MyMBGD functionality.

To cope with the explosive growth of microbial genome data owing to next generation sequencing technology, we need to improve the database construction procedure continuously. Since recently tens or even hundreds of genome sequences of the same species are available for many bacteria, we are trying to establish an efficient protocol to maintain all-against-all similarity data by separating intra-species comparisons and inter-species comparisons.

II. Orthologous gene classification 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 only 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 domain-level classification using the sum total of domain-level alignment scores. We developed a refinement pipeline to improve domain-level clustering, DomRefine, by optimizing the DSP score. DomRefine is now used to construct the standard ortholog table covering all the representative genomes stored in MBGD.

Domain-level classification is a unique feature of our ortholog classification system in MBGD. Now, we are analyzing the database and trying to characterize domain fusion events that occurred during prokaryotic and eukaryotic evolution.

III. Identification of the core structure conserved among taxonomically 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 various bacterial taxa to define their core gene sets.

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. In particular, we are trying to apply the RECOG system to comparative analyses of transcriptomic data and even metagenomic data.

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

Orthology is a key to integrate knowledge about various organisms through comparative analysis. 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). 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. On the basis of this framework, we are trying to integrate various kinds of microbial data using the ortholog information as a hub, as part of the MicrobeDB.jp project developed under the National Bioscience Database Center.

In addition, to facilitate the utilization of the RDF databases distributed worldwide, we developed a command-line tool, named SPANG, that simplifies querying distributed RDF stores using the SPARQL query language.

VI. Characterization of the gene repertoire of *H. pylori* pan-genome

Genomes of bacterial species can show great variation in their gene content, and thus systematic analysis of the entire gene repertoire, termed the "pan-genome", is important for understanding bacterial intra-species diversity. 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 (FindMobile) to define mobility of genes against the reference coordinate determined by the core alignment created by CoreAligner, and classified each non-core 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 (Figure 1). 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.

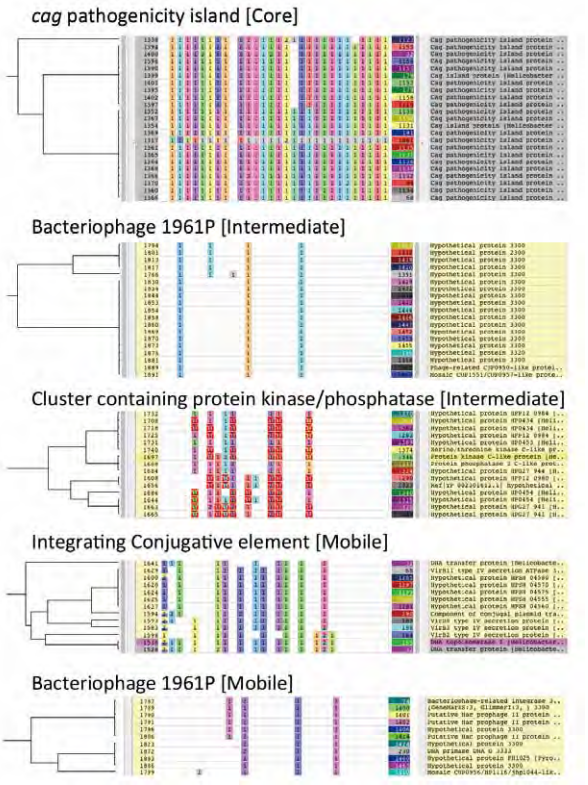

Figure 1. The five largest co-occurring gene clusters identified among 30 strains of *H. pylori* using the RECOG system

Publication List:

〔Original papers〕

- Abe, R., Toyota, K., Miyakawa, H., Watanabe, H., Oka, T., Miyagawa, S., Nishide, H., Uchiyama, I., Tollefsen, K.E., Iguchi, T., and Tatarazako, N. (2015). Diofenolan induces male offspring production through binding to the juvenile hormone receptor in *Daphnia magna*. Aquat. Toxicol. *159*, 44-61.
- Chiba, H., Nishide, H., and Uchiyama, I. (2015). Construction of an ortholog database using the Semantic Web technology for integrative analysis of genomic data. PLoS One *10*, e0122802.
- Fernández-Breis, J.T., Legaz-García, M.C., Chiba, H., and Uchiyama, I. (2015). Towards the semantic standardization of orthology content. Proc. Semant. Web Appl. Tool. Life Sci. 74-83.
- Hayashi, S., Kawaguchi, A., Uchiyama, I., Kawasumi-Kita, A., Kobayashi, T., Nishide, H., Tsutsumi, R., Tsuru, K., Inoue, T., Ogino, H., Agata, K., Tamura, K., and Yokoyama, H. (2015). Epigenetic modification maintains intrinsic limb-cell identity in *Xenopus* limb bud regeneration. Dev. Biol. *406,* 271-282.
- Kawai, M., Uchiyama, I., Takami, H., and Inagaki, F. (2015). Low frequency of endospore-specific genes in subseafloor sedimentary metagenomes. Environ. Microbiol. Rep. *7*, 341-350.
- Takami, H., Arai, W., Takemoto, K., Uchiyama, I., and Taniguchi, T. (2015). Functional classification of uncultured "*Candidatus* Caldiarchaeum subterraneum" using the Maple system. PLoS One *10*, e0132994.
- Uchiyama, I., Mihara, M., Nishide, H., and Chiba, H. (2015). MBGD update 2015: microbial genome database for flexible ortholog analysis utilizing a diverse set of genomic data. Nucleic Acids Res. *43*, D270-D276.