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. To efficiently explore the diversity of the microbial genomic data, MBGD pre-calculates ortholog tables for each major taxonomic group in the Taxonomy database, and provides several pages to display the entire picture of each pre-calculated ortholog table. 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). In addition, MBGD provides MyMBGD mode, which allows users to add their own genomes to MBGD.

The database continues to grow and now contains more than 2500 published genomes including 41 eukaryotic microbes and 4 multicellular organisms. To further enhance the database, we are now preparing to incorporate genomic data released as draft sequence data, which are now growing as rapid as complete sequences. Data will be incorporated in some pre-computed ortholog tables, and also will be provided for users to incorporate in their analysis through the MyMBGD function.

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 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 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 (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. This approach allows us to conduct further large-scale ortholog analysis including draft genome sequences. We are also extending the algorithm for handling metagenomic data. To infer the taxonomic position of the source organism 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.

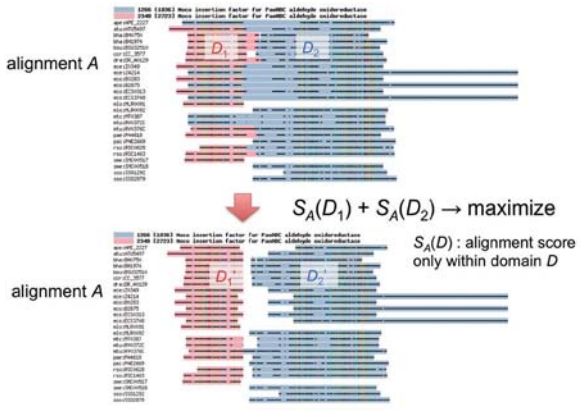

Figure 1. Definition of DSP score to evaluate domain-based classification in DomRefine.

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 (called "syntenic core"). To find such core structures, we developed a method named CoreAligner.

We systematically applied the method to bacterial taxa (family, genus and species) that contain a sufficient number of completed genomes stored in MBGD, and the resulting syntenic core gene sets were compared with corresponding universal core gene sets based on the conventional definition. As a result, syntenic core is generally larger than universal core, and typically the number of syntenic core genes is more stable than universal core when the number of genomes in the given taxa increases.

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 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, 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 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. Utilizing ortholog data to integrate microbial database using Semantic Web technologies

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. Toward this goal, we are trying to integrate various types of microbial data with genome/metagenome data and orthology relation using Semantic Web technologies. For this purpose, we described orthology relation and related data stored in MBGD and other databases using Resource Description Framework (RDF) and launched a SPARQL endpoint to query the database via the SPARQL language. 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. We identified co-occurring gene clusters using phylogenetic pattern clustering combined with neighboring gene clustering implemented in the RECOG system. In addition, we developed a method to define mobility of genes against the reference coordinate determined by the syntenic core alignment created by CoreAligner, and classified each accessory gene into mobility classes. On the basis of these analyses, we characterized the repertoire of accessory genes in *H. pylori* strains in terms of co-occurring gene clusters and mobility. This work is in collaboration with Prof. Kobayashi, Univ. Tokyo.

Publication List

〔Original papers〕

- Uchiyama, I., Mihara, M., Nishide, H., and Chiba, H. (2013). MBGD update 2013: the microbial genome database for exploring the diversity of microbial world. Nucleic Acids Res. *41*, D631-D635.
- Yahara, K., Furuta, Y., Oshima, K., Yoshida, T., Azuma, T., Hattori, M., Uchiyama, I., and Kobayashi, I. (2013). Chromosome painting in silico in a bacterial species reveals fine population structure. Mol. Biol. Evol. *30*, 1454-1464.

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

- 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. Molecular impact of juvenile hormone agonists on neonatal *Daphnia magna*. J. Appl. Toxicol. 2013 Sep. 5.