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 III 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. 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

II. Hierarchical strategy for creating ortholog tables

Previously, MBGD provided two types of ortholog tables: the standard ortholog table containing one representative genome from each genus covering the entire taxonomic range, and the taxon specific ortholog tables containing the genomes belonging to each taxonomic group (species, genus, family and so on). Although, by this approach, most of the "core genes", i.e. genes well conserved within species or genus, can be incorporated into the standard ortholog table, the genes that are not conserved in the representative genomes cannot be incorporated. Considering the great diversity of microbial species/genus genomes, a considerable amount of information may be lost in the current standard ortholog table. To address this problem, we developed a stepwise protocol to construct orthologous relationships. In this approach, we first create a within-species ortholog table for each species and construct a species level pan-genome by picking one representative gene from each orthologous group (Figure. 1). Next, we create a within-genus ortholog table for each genus using as input the species-level pan-genomes generated in the previous step and construct a genus level pan-genome. Finally, we create an ortholog table covering the entire taxonomic range by comparison of the genus level pan-genomes. We used a rapid similarity search program, UBLAST, to calculate all-against-all similarities for within-species and within-genus comparisons, which can drastically reduce the computation cost. By this approach, we can integrate various pan-genomes into a single ortholog table, which enables us to analyze evolutionary processes that generate within-species or within-genus diversity of microbes.

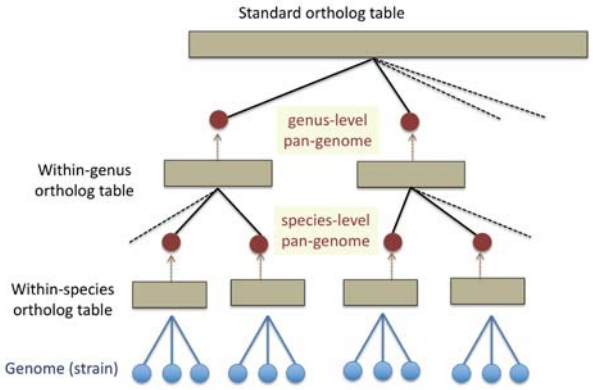

Figure 1. Hierarchical protocol for creating the standard ortholog table

III. 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.

We also developed a procedure to refine the DomClust classification based on multiple sequence alignments instead of pairwise 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 fact, it is different from conventional domain databases like Pfam in that it is based on orthology instead of homology. Particularly, this data is considered to be suitable for analyzing domain fusion events that occurred during evolution. Now, by analyzing the domain-level ortholog grouping data combined with taxonomic and functional information, we are trying to elucidate when and in what kind of genes domain fusion events frequently occurred.

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 (including *H. pylori* genome comparison described in Section VI below). In particular, in addition to microbial genome comparison, we are trying to apply RECOG to comparative analyses of transcriptomic data and 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 developed an ortholog ontology (OrthO) and described the ortholog information in MBGD in the form of the Resource Description Framework (RDF). To further standardize the ontology, we developed the Orthology Ontology (ORTH) in collaboration with Dr. Fernandez-Breis (Univ. Murcia) by integrating OrthO and OGO (another ortholog ontology developed by Dr. Fernandez-Breis) and reusing other existing ontologies.

On the basis of this framework, we have integrated various kinds of microbial data using the ortholog information as a hub, as part of the MicrobeDB.jp project (http://microbedb.jp/) 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. SPANG simplifies querying distributed RDF stores using the SPARQL query language, and provides a framework for reusing and sharing queries across the Web, thereby reducing the burden of writing complex queries in SPARQL.

VI. A novel approach for identification of genomic islands

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. 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. In addition, by clustering the accessory OGs on the basis of phylogenetic pattern similarity and chromosomal proximity, we identified 60 co-occurring gene clusters (CGCs). We are now trying to generalize this approach to identify genomic islands in various bacterial species.

Publication List:

〔Original papers〕

- Chiba, H., and Uchiyama, I. (2017). SPANG: A SPARQL client supporting generation and reuse of queries for distributed RDF databases. BMC Bioinformatics *18*, 93.
- Hayatsu, M., Tago, K., Uchiyama, I., Toyota, A., Wang, Y., Shimomura, Y., Okubo, T., Kurisu, F., Hirono, Y., Nonaka, K., Akiyama, H., and Takami, H. (2017). An acid-tolerant ammonia-oxidizing γ-proteobacterium from soil. ISME J. *11*, 1130-1141.
- Ikeda, T., Uchiyama, I., Iwasaki, M., Sasaki, T., Nakagawa, M., Okita, K., and Masui, S. (2017). Artificial acceleration of mammalian cell reprogramming by bacterial proteins. Genes Cells *22*, 918-928.
- Takami, H., Toyoda, A., Uchiyama, I., Itoh, T., Takaki, Y., Arai, W., Nishi, S., Kawai, M., and Ikeda, H. (2017). Complete genome sequence and expression profile of the commercial lytic enzyme producer *Lysobacter enzymogenes* M497-1, DNA Res. *24*, 169-177.

〔Review Article〕

- Uchiyama, I. (2017). Ortholog identification and comparative analysis of microbial genomes using MBGD and RECOG. In Protein function prediction, D. Kihara ed., Humana press, pp. 147-168.