## LABORATORY OF GENOME INFORMATICS

*Assistant Professor*
UCHIYAMA, Ikuo

*Postdoctoral Fellow:* *KATO, Masaki*

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 it, to understand complex living systems by integrating the data with current biological knowledge via the use of 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 interpret 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 by comparing 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, 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. By the application 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 viewing modes to display the entire picture of each ortholog table. For some closely related taxa, MBGD provides conserved synteny information calculated using the CoreAligner program. MBGD additionally provides MyMBGD mode, which allows users to add their own genomes to MBGD.

This year, we released the new version of MBGD, which contains 6318 genomes, including 5861 bacteria, 254 archaea, and 203 eukaryota. In addition to updating the database, we modified the protocol to construct ortholog table, which will be described in the next section.

### II. Hierarchical strategy for creating ortholog tables

MBGD previously calculated all-against-all similarities among the stored genomes and created two types of ortholog tables independently: 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).

The problem with this approach is twofold. Firstly, rapid accumulation of the genomic data from the same or closely related species substantially expands the size of all-against-all similarity data, while the increased net amount of information (*i.e.* the size of gene repertoire) is more limited. Secondly, the standard ortholog table contains only genes that are contained in the representative genomes, and thus a considerable amount of information may be lost from the standard ortholog table, if one were to consider within-species and within-genus genomic diversity.

To address these problems, we developed a stepwise protocol to construct orthologous relationships. First, for each species having at least two genomes, all-against-all similarities among the genomes belonging to that species are calculated and a within-species ortholog table is created. The species-level pan-genome is then created by picking one representative gene from each orthologous group. Next, for each genus having at least two species, all-against-all similarities among the species-level pan-genomes and a within-genus ortholog table is created. The genus-level pan-genome is then created by picking one representative gene from each orthologous group. Finally, all-against-all similarities among the genus-level pan-genomes are calculated and the standard ortholog table covering the entire taxonomic range is created. To calculate within-species or within-genus all-against-all similarities, we used a faster but less sensitive similarity search program, UBLAST. In this way, we can reduce the computation time required for all-against-all similarities.

An example of a hierarchical ortholog group is shown below (Figure 1), where an ortholog group containing Shiga toxins is shown. Analysis of such sporadically distributed genes was often not possible in the previous version of MBGD, because such an ortholog group was often not contained in the standard ortholog table.

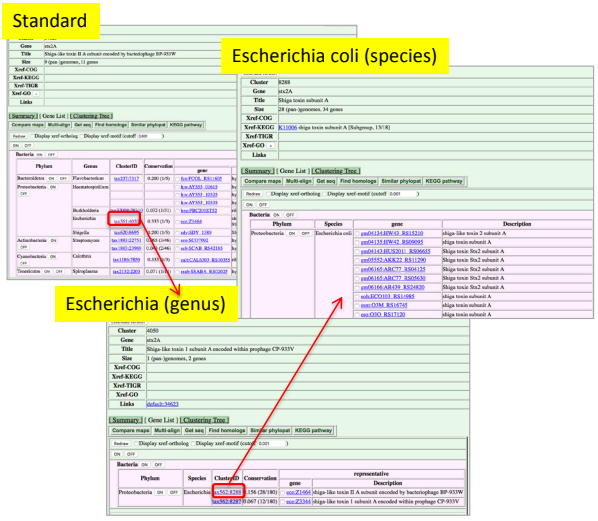


Figure 1. An example of a hierarchical ortholog group. Shown is the ortholog group containing Shiga-like toxins subunit A.

## 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 allows us to simultaneously compare many genomes. This 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. In particular, this data is considered to be suitable for analyzing domain fusion events that have occurred during evolution. By analyzing the domain-level ortholog grouping data combined with taxonomic and functional information, we are now 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 large-scale ortholog tables. The ortholog table viewer is a spreadsheet like viewer that can display an 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 performed 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 are continuing to develop the system and apply it to various genome comparison studies as part of various collaborative research projects (including *H. pylori* genome comparison described in Section VI below). In addition to microbial genome comparison, we are trying to apply RECOG to the 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 integrating knowledge about various organisms through comparative analysis. We have constructed an ortholog database using Semantic Web technology, and are aiming to integrate 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 auspices of the National Bioscience Database Center.

## 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. As we have already developed a procedure (coreAligner) to define the core genome as the genes conserved among the genomes of the given species, characterizing the remaining part of the genomes (non-core genomes) should be important for understanding the species’ diversity. To this end, we developed a method (FindMobile) to define mobility of genes against the reference coordinate determined by the core genome alignment, and classified each non-core gene into mobility classes. Combining this with a naive clustering procedure on the basis of phylogenetic pattern similarity and chromosomal proximity implemented in RECOG, we were able to identify genomic island candidates among the genomes of 30 *Helicobacter pylori* strains. We are now trying to generalize this approach to identify genomic islands in various bacterial species.

### Publication List:

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


- Uchiyama, I., Mihara, M., Nishide, H. , Chiba, H., and Kato, M. MBGD update 2018: microbial genome database based on hierarchical orhtology relations covering closely related and distantly related comparisons. Nucleic Acids Res. 2018 Nov 20.