LABORATORY OF GENOME INFORMATICS

Research Associate: *UCHIYAMA, Ikuo*

The accumulation of biological data has recently been accelerated by various high-throughput so-called “omics” technologies such as genomics, transcriptomics, proteomics and so on. The field of genome informatics is aimed at utilizing this data, or finding some 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 or tools for comparative genome analysis, which is a useful approach for finding functional or evolutionary clues to interpreting genomic information of various species. The current focus of our research topics is on comparative analysis of microbial genomes, the number of which is now beyond a hundred, as a basic model system for understanding the variety of life through the comparative analysis of numerous genomic sequences simultaneously.

I. Construction of microbial genome database for comparative analysis

The number of completed microbial genome sequences is growing rapidly, and nearly three hundred genome sequences of various levels of relatedness have already been made available. The role of comparative genomics becomes much more important in order to utilize this large number of sequences not only for elucidating commonality in all of life, but also for understanding the evolutionary diversity within various groups, as well as for understanding the evolutionary processes or mechanisms producing such diversity.

We have been developing and maintaining a database system for comparative analysis of microbial genomes named MBGD (http://mbgd.genome.ad.jp/) (Figure 1). The key components of MBGD include i) an efficient algorithm that can classify genes into orthologous groups using precomputed all-against-all similarity search results (see below), ii) a user interface that is designed for users to explore the resulting classification in detail, and iii) an incremental updating process for similarities between genes and other data, which enables the system to provide the latest data rapidly. By this approach, MBGD is now the world’s largest database of its kind. Moreover, by specifying a set of organisms, users can obtain the appropriate classification results that they require using the latest data available.

The constructed classification table can be used for comparative analyses from various points of view, such as phylogenetic pattern analysis, gene order comparison and detailed gene structure comparison (Figure 1).

II. Hierarchical clustering algorithm for constructing orthologous groups of multiple genomes

Although ortholog identification is a crucial first step in comparative genomics, a scheme for large-scale ortholog grouping is yet to be established. The conventional approach to the identification of orthologs between two genomes is the so-called bidirectional best-hit (BBH) criterion, where two genes, *a* and *b*, in the genomes *A* and *B*, respectively, are considered to be orthologs if *a* is the best hit of *b* in genome *A* and vice versa. The Clusters of Orthologous Groups (COGs) Database, a widely used curated database for ortholog grouping, was constructed using this approach, although the overall construction process has included additional complex procedures such

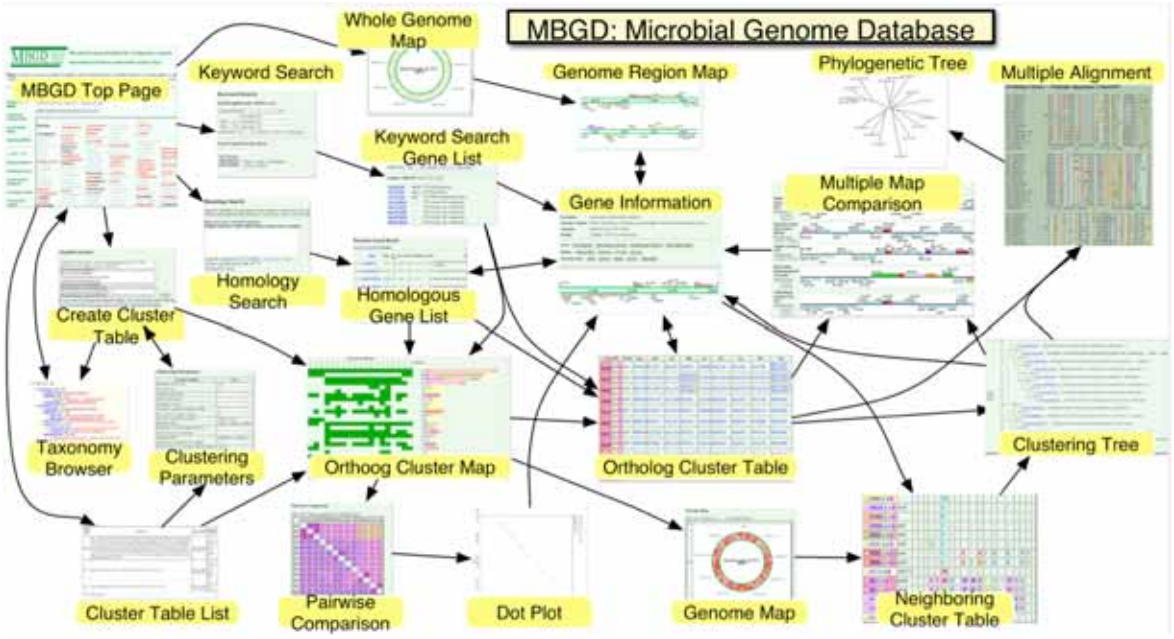

Figure 1. A flow diagram of the MBGD user interface

as the addition of species-specific paralogs, the splitting of proteins into multiple domains if required, and other case-by-case manual modifications

As a part of the core technologies of the MBGD system, we have developed a rapid automated method of ortholog grouping which is effective enough to allow the comparison of hundreds of genomes simultaneously. The method takes as input all-against-all similarity data and classifies genes based on the traditional hierarchical clustering algorithm UPGMA. In the course of clustering, the method detects domain fusion or fission events, and splits clusters into domains if required. The subsequent procedure splits the resulting trees such that intra-species paralogous genes are divided into different groups so as to create plausible orthologous groups. As a result, the procedure can split genes into the domains minimally required for ortholog grouping.

The procedure, named DomClust, was tested using the COG database as a reference. When comparing several clustering algorithms combined with the conventional BBH criterion, we found that our method generally showed better agreement with the COG classification. By comparing the clustering results generated from datasets of different releases, we also found that our method showed relatively good stability in comparison to the BBH-based methods.

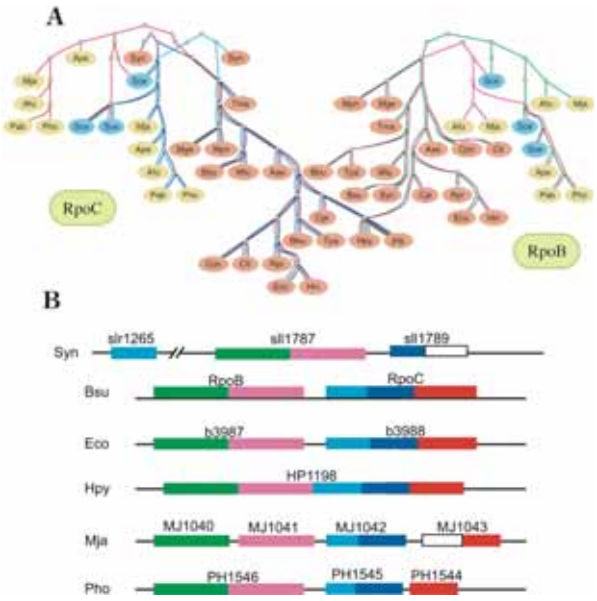

Figure 2. Orthologous groups of RNA polymerase beta (RpoB) and beta' (RpoC) subunits, as an example of DomClust classification. A) Hierarchical clustering trees constructed by the DomClust program. Each domain is drawn in a different color. An abbreviated species name (taken from the COG database) is shown on each leaf, which is colored according to the kingdom: salmon, bacteria; khaki, archaea; sky-blue, eukaryotes. B) Schematic illustration of the gene structures of RpoB and RpoC in selected genomes.

III. Identification of the common core structure of phylogenetically related genomes

It is known that horizontal transfer as well as vertical transfer has played important roles in prokaryotic evolution. Because of this complexity, further investigation is required in order to obtain a clearer picture of the bacterial genome evolution. Extensive comparison of multiple genomes that are closely or moderately related to each other should give many clues to understanding evolutionary processes. Such data is now rapidly accumulating in our MBGD database.

We are trying to identify a common "core structure" of phylogenetically related genomes, which is defined as a set of sufficiently long consecutive genomic segments in which gene orders are conserved among multiple genomes so that they are likely to have been inherited from the common ancestor mainly through vertical transfer. For this purpose, we have developed a graph-based algorithm for aligning conserved regions of multiple genomes by ordering orthologous groups so as to retain the conserved gene orders as much as possible.

The method was applied to the comparison of *Bacillus*-related species whose genome sequences have been determined including alkaliphilic *B. halodurans*, halotolerant *Oceanobacillus iheyensis* and thermophilic *Geobacillus kaustophilus*, in addition to well-known laboratory strain *B. subtilis* and pathogenic *B. anthracis* and *B. cereus*. These organisms – except for *B. anthracis* and *B. cereus* – are moderately diverged each other and belong to distinct major clusters in the 16S rRNA phylogenetic tree. Overall genomic structures are primarily well conserved between them, which can be confirmed by dotplot analyses where large collinear regions along the diagonal lines can easily be seen.

Using orthologous groups of *Bacillus*-related species with *Staphylococcus aureus* as an outgroup generated by the DomClust program, we constructed genome alignments by the above algorithm. From this alignment, we were able to identify the common core structure of *Bacillus* genomes comprising about 1500 genes. It appears that most of the important genes are included in the resulting core gene set. Indeed, the set contains most of 271 *B. subtilis* essential genes that were primarily determined by a systematic inactivation experiment. Further investigation of the core gene set revealed characteristic distributions of function categories in the core and non-core gene sets.

Publication List:

Original paper

Ishikawa, K., Watanabe, M., Kuroita, T., Uchiyama, I., Bujnicki, J.M., Kawakami, B., Tanokura, M., and Kobayashi, I. (2005). Discovery of a novel restriction endonuclease by genome comparison and application of a wheat-germ-based cell-free translation assay: PabI (5'-GTA/C) from the hyperthermophilic archaeon *Pyrococcus abyssi*. Nucleic Acids Res. *33*, e112.