LABORATORY OF GENOME INFORMATICS

Assistant Professor
UCHIYAMA, Ikuo

The accumulation of biological data has recently been accelerated by various high-throughput "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 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. Extracting useful information from such a growing number of genomes is a major challenge in genomics research. Interestingly, many of the completed genomic sequences are closely related to each other. We are now trying to develop methods and tools to conduct comparative analyses not only of distantly related genomes but also of closely related genomes, since we can extract different types of information about biological functions and evolutionary processes from comparisons of genomes at different evolutionary distances.

Ⅰ. Microbial genome database for comparative analysis

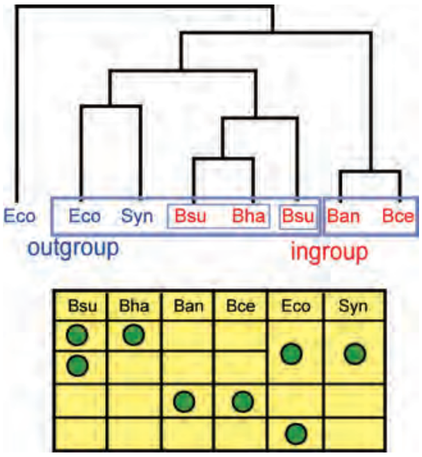
The microbial genome database for comparative analysis (MBGD) is a comprehensive platform for microbial comparative genomics. The central function of MBGD is to create orthologous groups among multiple genomes from precomputed all-against-all similarity relationships using the DomClust algorithm (see Section II below). By this algorithm, MBGD not only provides the 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. The latter feature is especially useful when the user's interest is focused on some taxonomically related organisms. 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. For researchers who are interested in ongoing genome projects, MBGD also provides a service called "My MBGD," which allows users to add their own genome sequences to MBGD for the purpose of identifying orthologs among both the new and the existing genomes.

The database now contains more than 500 published genomes and the number continues to grow. MBGD is available at http://mbgd.genome.ad.jp/.

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

As part of the core technologies of the MBGD system we have developed a rapid automated method of ortholog grouping, named DomClust, 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 in such a way 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.

As the number of completed genomic sequences grows, comparison among closely related as well as distantly related genomes has become more important for understanding the function and evolution of various genomes. For evolutionary analysis of a target set of related organisms (ingroup), we often need another set of organisms outside of that group (outgroup) to correctly infer evolutionary processes. To incorporate this concept into the ortholog analysis, we have enhanced the DomClust algorithm to impose constraints on the resulting orthologous groups such that the outgroup species should come outside of the ingroup species (Figure 1). The resulting table has a nested structure when a duplication event occurs within the ingroup lineage.



Bsu	Bha	Ban	Bce	Eco	Syn
●	●			●	●
●					
		●	●		
				●	

Figure 1. Converting from a gene tree (top) to a nested ortholog table (bottom)

Ⅲ. Identification of core structures conserved among moderately related microbial genomes

A growing body of evidence supports the theory that both horizontal transfer and vertical transfer have played significant roles in prokaryotic evolution. Despite the complexity of evolution suggested from these observations, it can be argued that prokaryotic phylogeny can still be inferred using a certain subset of genes ("core genes") that have mainly transferred vertically throughout the evolutionary process. We are trying to identify a common "core structure" of 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 a common ancestor mainly through vertical transfer. For this purpose, we have developed a graph-based algorithm for aligning conserved regions of multiple genomes. The algorithm finds an order of pre-identified orthologous groups so as to retain, as much as possible, the conserved gene orders (Figure 2).

The method was applied to genome comparisons of the families *Bacillaceae* and *Enterobacteriaceae*. Using orthologous groups generated by the DomClust program, we constructed genome alignments and identified common core structures comprising about 1400 genes for Bacillaceae and 1900 genes for Enterobacteriaceae. Despite the difference in the overall proportions of the core genes between these datasets, the proportion of the core genes in each functional category primarily exhibit a similar tendency: functional categories related to primary metabolism, genetic information processing, and cellular processes generally contain a higher proportion of core genes, while categories of membrane transport, signal transduction and secondary metabolism contain a lower proportion of core genes. In addition, it also turned out that these core structures contain most of the essential genes identified in *Bacillus subtilis* and *Escherichia coli*, respectively.

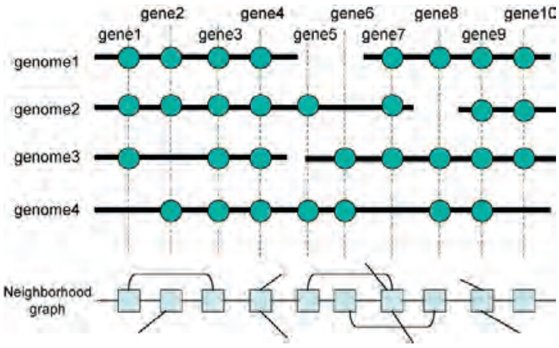


Figure 2. Construction of core genome alignment based on gene order conservation

Ⅳ. Research environment for comparative genomics

We have been developing a new system named RECOG (Research Environment for COmparative Genomics), which aims to extend the current MBGD system by incorporating the above-described approaches including the ingroup/outgroup distinction in ortholog grouping and the core structure extraction among related genomes. In addition, the RECOG system has several advanced features that allow users to perform more flexible phylogenetic pattern analyses.

The entire RECOG system employs client-server architecture. The RECOG server program has been developed based on the MBGD server and contains database construction protocol in MBGD including all-against-all similarity search calculation, as well as the extended version of the DomClust program. Unlike MBGD, however, users can install the RECOG server on their local machines to analyze their own genomic data. Alternatively, users can connect to the public RECOG server to analyze publicly available data.

The RECOG client program is a Java application that runs on a local machine by receiving data from any available RECOG server. The main window of the RECOG client consists of three parts: taxonomic tree viewer (left), ortholog table/phylogenetic pattern map viewer (center) and gene information viewer (right) (Figure 3). In the taxonomic tree viewer, users can specify either a set of organisms to be analyzed (as ingroup or outgroup), or conditions to filter phylogenetic patterns (i.e. presence or absence of each gene in each genome) to be displayed. The ortholog table viewer displays the entire ortholog table. By semantic zooming functionality, users can see from the entire picture of the phylogenetic pattern map to the full details of the ortholog table. In the gene information viewer, users can see detailed information about specified orthologous groups and genes belonging to each of the groups. Several sorting and filtering functions have been implemented for modifying the display of the ortholog table viewer, including sorting by functional category and gene name or by gene order of a specified genome, and filtering by keywords or by phylogenetic pattern conditions.

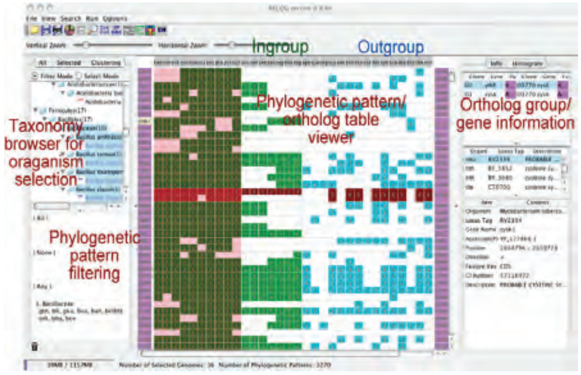

Ingroup
Outgroup
Taxonomy browser for oraganism selection
Phylogenetic pattern/ ortholog table viewer
Ortholog group/ gene information
Phylogenetic pattern filtering

Figure 3. A snapshot of the main window of the RECOG system

Publication List

〔Original paper〕

- Uchiyama, I. (2007). MBGD: a platform for microbial comparative genomics based on the automated construction of orthologous groups. Nucleic Acids Res. *35*, D343-D346.