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 developing 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 III 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 around 700 published genomes including 13 eukaryotic microbes (fungi and protozoa) and *C. elegans* as a reference. MBGD is available at http://mbgd.genome.ad.jp/.

Ⅱ. 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: a "core gene pool" that comprises intrinsic genes encoding the proteins of basic cellular functions, and a "flexible gene pool" that comprises HGT-acquired genes encoding proteins which function under particular conditions, such as genomic islands. Therefore, the identification of the set of intrinsically conserved genes, or the genomic core, among a taxonomic group is crucial not only for establishing the identity of each taxonomic group, but also for understanding prokaryotic diversity and evolution. Here, 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. We developed a method for aligning conserved regions of multiple genomes, which finds the order of pre-identified orthologous groups (OGs) that retains to the greatest possible extent the conserved gene orders.

The program, named CoreAligner, requires a set of well-conserved OGs. We compiled them using the MBGD server, and considered an OG "conserved" when it was present in at least half of the genomes. Next, a neighborhood graph was constructed using a set of conserved neighborhood pairs, which are defined as two conserved OGs that are located within 20 genes in at least half of the genomes. Our algorithm for constructing the alignments of the core genome structures is based on finding the longest path of the conserved neighborhood graph. A similar algorithm was previously developed mainly for identifying much shorter but more widely conserved gene clusters such as operons, but unlike that method, our method considers not only genes in the same direction but also those in the opposite direction as neighboring genes, and thereby generally generates longer alignments. In addition, our method uses the dynamic programming algorithm for calculating the longest path.

The method was applied to genome comparisons of two well-characterized families, *Bacillaceae* and *Enterobacteriaceae*, and identified their core structures comprising 1438 and 2125 OGs, respectively (Figure 1), which correspond to a third of the number of the B. *subtilis* genes (4105) and half of the E. *coli* genes (4237), respectively. The core sets contained most of the essential genes (90%) and their related genes, which were primarily included in the intersection of the two core sets comprising around 700 OGs. The definition of the genomic core based on gene order conservation was demonstrated to be more robust than the simpler approach based only on gene conservation. We also investigated the core structures in terms of G+C content homogeneity and phylogenetic congruence, and found that the core genes primarily exhibited the expected characteristic (i.e., being indigenous and sharing the same history) more than the non-core genes.

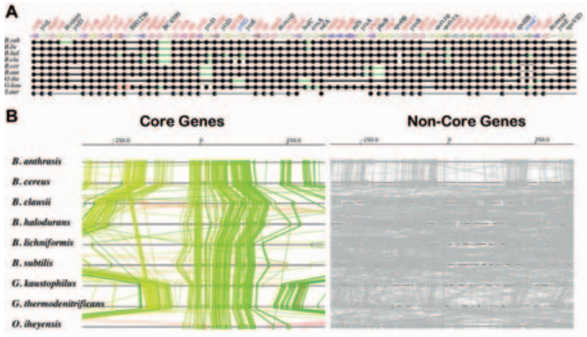

Figure 1. Core genome alignment. (A) Part of a schematic representation of genome alignment obtained from the *Bacillaceae* dataset. (B) Comparative genome map showing the locations of the core (left) and non-core (right) genes, where the latter are extensively crossed with each other.

Ⅲ. Enhancement of the algorithm for identifying orthologous groups among multiple genomes

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.

As the number of completed genomic sequences grows, comparison among closely related as well as distantly related genomes has become 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 so that the outgroup species should come outside of the ingroup species. The resulting table has a nested structure when a duplication event occurs within the ingroup lineage.

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 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 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). Users can choose a set of genomes to compare in the taxonomic tree viewer and run the DomClust program to identify orthologous groups among them. The result is displayed in the ortholg table viewer, where users can see the entire picture of the phylogenetic pattern map as well as the full details of the ortholog table using semantic zooming functionality. Several sorting and filtering functions have been implemented for modifying the display of the ortholog table viewer, including sorting by functional category or by the gene order of a specified genome, and filtering by keywords or by phylogenetic pattern conditions. The RECOG system also contains an interface for the CoreAligner program, by which users can see an alignment display and a comparative map display of conserved core regions identified by the CoreAligner program (Figure 2).

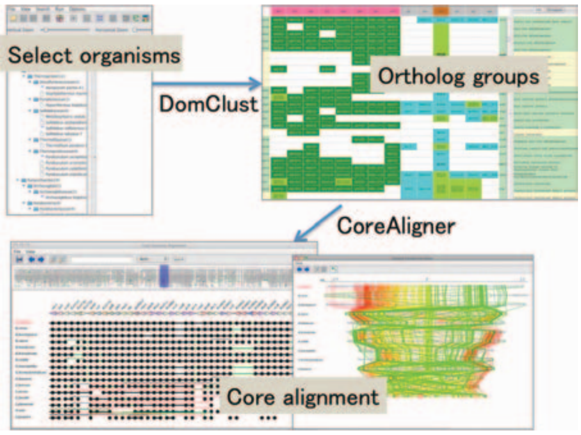

Figure 2. A typical usage of the RECOG system including DomClust and CoreAligner analyses.

Publication List

〔Original papers〕

- Uchiyama, I. (2008). Multiple genome alignment for identifying the core structure among moderately related microbial genomes. BMC Genomics, *9*, 515.
- Nakayama, K., Yamashita, A., Kurokawa, K., Morimoto, T., Ogawa, M., Fukuhara, M., Urakami, H., Ohnishi, M., Uchiyama, I., Ogura, Y., Ooka, T., Oshima, K., Tamura, A., Hattori, M., Hayashi, T. (2008). The Whole-genome sequencing of the obligate intracellular bacterium *Orientia tsutsugamushi* revealed massive gene amplification during reductive genome evolution. DNA Res. *15*, 185-199.