LABORATORY OF GENOME INFORMATICS

Research Associate: *UCHIYAMA, Ikuo*

The accumulation of biological data has recently been accelerated by various high-throughput ìomicsî 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 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 userís 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.

The database now contains more than 300 published genomes and the number continues to grow. For researchers who are interested in ongoing genome projects, we have started a new service called ìMy MBGD,î 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 (Figure 1). Furthermore, in order to make the rapidly accumulating information on closely related genome sequences available, we enhanced the interface for pairwise genome comparisons using the CGAT interface (see Section III below), which allows users to see nucleotide sequence alignments of non-coding as well as coding regions. MBGD is available at http://mbgd.genome.ad.jp/.

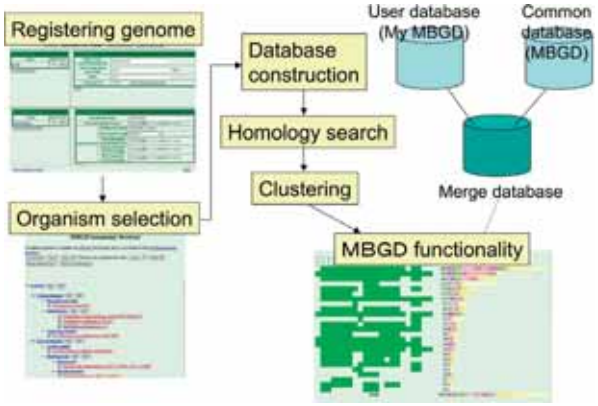

Figure 1. The My MBGD functionality, which allows users to add their own genome data to the MBGD database.

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 has yet to be established. In fact, the conventional approach to the identification of orthologs, called the bidirectional best-hit (BBH) criterion, is known to have several drawbacks, and the establishment of orthologous relationships in a database like the Clusters of Orthologous Groups (COGs) database requires additional complex procedures such 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, 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.

We are now trying to enhance the algorithm so as to take partial phylogenetic relationships into account to identify orthologous relationships among a set of closely related genomes with some outgroup species.

III. Comparative genome analysis tool for the analysis of complex evolutionary changes between closely related genomes

The recent accumulation of closely related genomic sequences provides a valuable resource for the elucidation of the evolutionary histories of various organisms. However, although numerous alignment calculation and visualization tools have been developed to date, the analysis of complex genomic changes, such as large insertions, deletions, inversions, translocations and duplications, still presents certain difficulties.

We have developed a comparative genome analysis tool, named CGAT, which allows detailed comparisons of closely related bacteria-sized genomes, mainly through visualizing middle-to-large-scale changes to infer the underlying mechanisms (Figure 2). CGAT displays precomputed pairwise genome alignments on both dotplot and alignment viewers with scrolling and zooming functions and allows users to move along the pre-identified orthologous alignments. Users can place several types of information on this alignment, such as the presence of tandem repeats or interspersed repetitive sequences and changes in G+C contents or codon usage bias, thereby facilitating the interpretation of the observed genomic changes. In addition to displaying precomputed alignments, the viewer can dynamically calculate the alignments between specified regions; this feature is especially useful for examining the alignment boundaries. Besides the alignment browser functionalities, CGAT also contains an alignment data construction module that provides a general framework for the calculation of genome-scale alignments using various existing programs as alignment engines, which allows users to compare the outputs of different alignment programs.

In collaborative studies with Dr. Kobayashiís group (Tokyo Univ.), we were able to conduct several comparative analyses using the earlier versions of this program to infer the evolutionary history of apparently complex genome changes between closely related eubacteria and archaea.

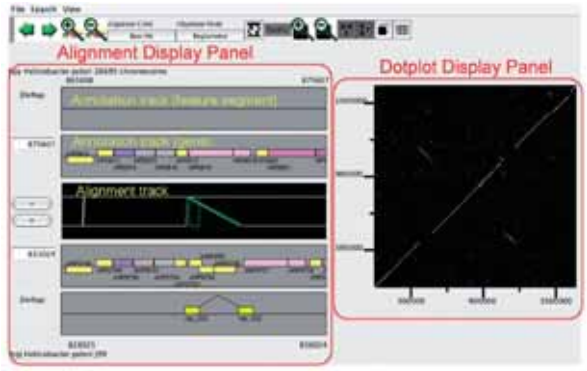

Figure 2. CGAT AlignmentViewer, which consists of an alignment display and a dotplot display. The example shown is ìinsertion with long target duplicationî, which was discovered during a comparison of two strains (26695 and J99) of *Helicobacter pylori* (Nobusato et al. 2000).

IV. Identification of core structures conserved among phylogenetically related genomes

It is known that both horizontal transfer and vertical transfer have 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 provide many clues for understanding evolutionary processes. Such data is now rapidly accumulating in our MBGD database.

We are trying to identify a common ìcore structureî 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 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.

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 1500 genes for *Bacillaceae* and 2000 genes for *Enterobacteriaceae*. It turned out that these core structures contain most of the essential genes identified in *Bacillus subtilis* and *Escherichia coli*, respectively. Further investigation for generalizing our approach is in progress.

Publication List:

Original papers

Kawai, M., Uchiyama, I., and Kobayashi, I. (2006). Genome comparison *in silico* in *Neisseria* suggests integration of filamentous bacteriophages by their own transposase. DNA Res. *12*, 389-401.

Kawai, M., Nakao, K., Uchiyama, I., and Kobayashi, I. (2006). How genomes rearrange: Genome comparison within bacteria *Neisseria* suggests roles for mobile elements in formation of complex genome polymorphisms. Gene *383*, 52-63.

Tsuru, T., Kawai, M., Mizutani-Ui, Y., Uchiyama, I., and Kobayashi, I. (2006). Evolution of paralogous genes: Reconstruction of genome rearrangements through comparison of multiple genomes within *Staphylococcus aureus*. Mol. Biol. Evol. *23*, 1269-85.

Uchiyama, I. (2006). Hierarchical clustering algorithm for comprehensive orthologous-domain classification in multiple genomes. Nucleic Acids Res. *34*, 647-658.

Uchiyama, I., Higuchi, T., and Kobayashi, I. (2006). CGAT: a comparative genome analysis tool for visualizing alignments in the analysis of complex evolutionary changes between closely related genomes. BMC Bioinformatics *7*, 472.