Microbial Genomics Pipeline for Comparative Studies

Sungsoo Kang, Sang-Yoon Kim, and Jong Bhak,

Korean BioInformation Center, KRIBB, Daejeon

Publically available microbial genome sequences are increasing rapidly owing to fast DNA sequencing technology [1, 2]. Automated bioinformatic pipelines to process them are indispensable for their comparative analyses. Such comparison is also critical for the identification of coding regions on microbial genomes [3].

We have developed a microbial genomics pipeline, which is useful for comparative genomics (Fig. 1). We compared various microbial genomes with all the theoretical proteomes of reference microbial organisms to obtain their best sequence alignments. We identified representative coding regions on each genome, and provided sequence identity profiles for each coding region across all the reference organisms. The representative coding regions were successfully used to define gene structure in bacteria and archaea [3]. They were also useful for the identification of authentic frame-shifts and sequencing errors in coding regions. By averaging sequence identities over all the representative coding regions, we measured pair-wise genomic similarities among reference organisms, and constructed the corresponding phylogenetic tree. Furthermore, by integrating various biological information, such as taxonomy, diseases, cellular functions, habitats, and phenotypes, we identified various lineage-specific genes and differentially-conserved genes, which explained biological characteristics of each microbial organism.

[1] GOLD: Genomes Online Database, http://www.genomesonline.org/.

[2] von Bubnoff A, *Cell*, **132**, 721 (2008).

[3] Kang S, Yang SJ, Kim S, and Bhak J, *Bioinformatics,* **23**, 3088 (2007).

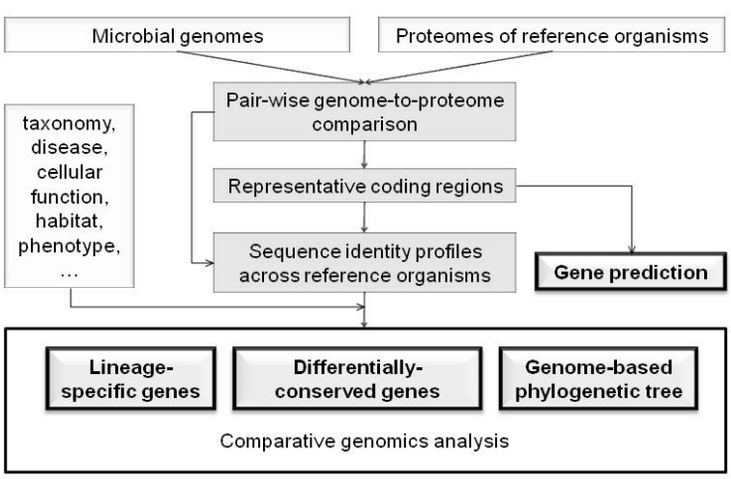

Figure 1. Overview of microbial genomics pipeline