Application of GGKbase to analyze "omic" data from acetate-stimulated subsurface microbial communities

Brian C. Thomas, Kelly C. Wrighton, Ken-ichi Ueda, Andrea Singh, **Jill Banfield***
University of California, Berkeley;

http://geomicrobiology.berkeley.edu/rifle/acd_ggkbase.html

Rifle IFRC project goal: To simultaneously analyze genomes reconstructed from community metagenomic data to provide functional insight

Cultivation-independent approaches provide access to the wide diversity of microorganisms in natural environments. Sequence data (metagenomic information) is foundational to most studies of natural microbial communities. In addition to providing insight into metabolic potential, these data enable functional analysis through proteomics (proteogenomics) and provide context for transcriptomic and metabolomics information.

Given the vast sizes of metagenomic datasets, extraction of biological and biogeochemical insight is challenging. Here, we present GGKbase as a means for data management and analysis, demonstrating its utility through application to microbial community datasets collected during the early, mid, and late phase of iron-reduction (secondary stimulation) at DOE's Rifle IFRC site. The Kbase is designed for analysis of well-assembled datasets, where much of the data have been assigned to near-complete genome bins. Although relatively uncommon at this time, such datasets will be the norm in the near future, as sequencing and bioinformatics methods improve.

GGkbase is a multi-user, list-based, social/sharing approach for analysis of the metabolism of individual organisms and comparative metabolic analysis at the community level. For an individual organism, is possible to browse the gene content of genome fragments, to investigate the functional predictions, and to modify them. The linked pages also document sequence similarity to other organisms, and provide access to proteomic information. Individual genes or groups of genes belonging to a pathway can be assigned to one of more lists, as determined by the investigator, and these lists can be shared with other users (new users can be invited to participate in curating a list). Because the lists are driven by a keyword search (or EC number, GO term etc.), genes can be identified and classified simultaneously across the entire dataset. This establishes metabolic profiles using tens, hundreds, and potentially thousands of genes at a time. It is also possible to summarize the gene and pathway capacities across the dataset using "genome summary" visualization. This allows recognition of characteristics shared by multiple organisms, thus is useful for investigating the molecular underpinnings of ecosystem metabolic processes. Application to the Rifle metagenomic dataset comprising enabled efficient analysis of 87 genomic datasets. Genome summaries provide a means to rapidly communicate information about genome completeness (based on marker gene content) and metabolic strategies, e.g., obligate fermentation. Development of GGKbase is supported by DOE's Knowledgebase Program, Susan Gregurick, Program Manager. Rifle IFRC research is supported by U.S. Department of Energy under Contract No. DE-AC02-05CH11231