

GGKbase, a portable knowledgebase for analysis and integration of “omic” data from microbial communities

Brian C. Thomas^{*1}, Ken-ichi Ueda¹, Rahul Basu², Chongle Pan³, Trent Northen³, Benjamin Bowen^{*2}, **Jill Banfield^{*1}**

¹*University of California, Berkeley;* ²*Lawrence Berkeley National Laboratory;* ³*Oak Ridge National Laboratory*

<http://genegrabber.berkeley.edu/>

Project Goals: To develop a knowledgebase for analysis and integration of ‘omics’ data from microbial communities

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 as it enables functional analysis through proteomics (proteogenomics) and provides context for transcriptomic and metabolomics information.

Given the vast dataset sizes, extraction of biological and biogeochemical insight from ‘omic’ datasets is challenging. In the current project, we have constructed a workflow and knowledgebase that enables recovery, efficient and effective display, manipulation, and interrogation of such information. Most important, the structure is portable. Developed initially for analysis of data from acid mine drainage microbial communities, the structure has already been populated by information from multiple other projects, including the DOE Rifle IFRC metagenomics and proteomics efforts.

The pipeline includes all components required for analysis of next-generation sequencing information, from assembly through binning and functional annotation. An explicit goal of GGKbase is the recovery of near-complete genomes, a feature that distinguishes our approach from others (e.g., MGRAST). Curated and binned genome fragments are grouped into organismal “bins” from which inferences about metabolic potential can be made. Genes for which proteins have been identified from one or a series of samples using the open reading frames predicted from the metagenomic data are flagged at the “genome browsing” level, and detailed information about abundance and distribution can be accessed, gene-by-gene.

We have developed GeneGrabber (a component of GGKbase), a multi-user, list-based, social/sharing approach for analysis of the metabolism of individual organisms and comparative metabolic analysis at the community level. 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, including the ability to invite new users 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.

Understanding an ecosystem's metabolic potential is a complex task. Leveraging the extensive, content-based lists created for each metagenomics resource, we developed a tool within GeneGrabber for visualizing the extent of metabolic machinery present in the data. This visualization, termed "genome summary," is invaluable for exploring metabolic pathways and can identify which organisms in the community are responsible for a process. The genome summary is a useful tool for investigating the molecular underpinnings of ecosystem metabolic processes.

The GGKbase system has been engineered to access other 'omics' data resources, without having to resort to database federation. GGKbase uses representational state transfer (REST) to tap into other information sources and we have developed a caching system using Redis to accelerate user-centric data access. We have also developed an API to access the GGKbase resource both in Ruby as well as just using simple URL access. Currently, GGKbase includes both metagenomic and proteomic data. Additionally, we have expanded the GGKbase structure to now include metabolomics data and have developed a system for processing and displaying this data called MetaboliteAtlas. MetaboliteAtlas, like all components in the GGKbase utilizes a RESTful architecture and includes a separate front-end web display for in-depth metabolomics investigations.

GGKbase is under continual development, currently focused around addition of metabolomic and transcriptomic data. As more researchers begin using GGKbase, new ideas are captured into the structure. Current challenges in increasing the scale of the community metagenomics approach include automation of time-intensive binning steps and aspects of time series data analysis.